Regularized sparse parameter identification for stochastic systems with collinearity

Jian Guo, Ying Wang, *Member, IEEE*, Yanlong Zhao, *Senior Member, IEEE*, Ji-Feng Zhang, *Fellow, IEEE*

***Abstract*— It is well known that sparse identification in high-dimensional engineering systems faces two major obstacles: (i) ill-conditioning caused by strong collinearity among variables, and (ii) feedback-induced endogeneity that produces non-independent and identically distributed (i.i.d.) and non-stationary observations. While these issues severely hinder reliable model discovery, existing methods typically address them in isolation. In contrast, this paper jointly considers these obstacles and proposes a two-stage adaptive identification scheme that combines a weighted L_1 penalty for sparsity and exact support recovery with a quadratic penalty that mitigates ill-conditioning from highly correlated regressors. Under excitation requirements much weaker than classical persistent excitation, global guarantees are derived via martingale techniques: all zero coefficients of the sparse parameter vector are identified with probability one after finitely many observations, the estimates of the nonzero coefficients converge almost surely to their true values, and these estimates are asymptotically normal. A group-selection property is also established: highly correlated yet relevant variables are retained together, avoiding the typical L_1 limitation under collinearity. The results do not rely on i.i.d. assumptions and are applicable to the identification of closed-loop linear stochastic systems with adaptive regulation. Numerical examples and a real robot-arm case study corroborate the theory and demonstrate performance improvements.**

***Index Terms*— Stochastic system, sparse identification, strong consistency, asymptotic normality, collinearity.**

The research was supported by the National Natural Science Foundation of China under Grants 62433020, 62025306, 62303452, and T2293770, CAS Project for Young Scientists in Basic Research under Grant YSBR-008, and China Postdoctoral Science Foundation under Grant 2022M720159. (Corresponding author: Ji-Feng Zhang.)

Jian Guo is with the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: j.guo@amss.ac.cn).

Ying Wang is with the Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm 11428, Sweden, and the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. (e-mail: wangying96@amss.ac.cn).

Yanlong Zhao is with the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: ylzhao@amss.ac.cn).

Ji-Feng Zhang is with the School of Automation and Electrical Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China, the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China (jif@iss.ac.cn).

I. Introduction

With the rapid growth in the size and complexity of datasets, the need for efficient data processing and analysis has become increasingly urgent. Sparse identification offers a powerful way to manage such complexity by extracting only the most relevant information [1]. It has attracted considerable attention in a wide range of disciplines, including systems and control [2], signal processing [3], statistics [4], and machine learning [5]. By aiming to accurately recover both zero and nonzero elements of sparse parameter vectors, sparse identification represents complex data with a minimal set of features, producing parsimonious and reliable predictive models [6].

In statistics, a wide range of sparsity-promoting techniques has been developed to perform variable selection by penalizing coefficient magnitudes, thereby retaining only the most relevant predictors. Representative examples include the Least Absolute Shrinkage and Selection Operator (LASSO) [7], Smoothly Clipped Absolute Deviation (SCAD) [8], elastic net [9], adaptive LASSO [10], and Minimax Concave Penalty (MCP) [11]. Within the time series domain, sparse regression has been extensively investigated in the literature. For example, [12] proposed regularized methods for high-dimensional vector autoregressive (VAR) models; [13] applied adaptive LASSO for large VAR modeling; and [14] examined the finite-sample performance of regularized autoregressive estimators. More recently, [15] reviewed how modern machine learning techniques can advance time series forecasting, with emphasis on methodological developments and applications. While these works explicitly consider temporal dependence, they usually rely on assumptions such as stationarity of the noise and weak dependence. In the systems and control community, sparse adaptive algorithms have also been explored. Notable contributions include the recursive L_1-regularized least-squares (LS) algorithm proposed by [16] and the adaptive greedy algorithm in [17] for online sparse recovery, with applications to FIR channel identification and linear prediction. Despite their effectiveness in specific scenarios, these existing statistical and control approaches typically rest on restrictive assumptions, such as independent and identically distributed (i.i.d.) stochastic inputs, known innovation distributions, or specific dependence structures, and may not offer general theoretical guarantees [18]. More recently, [19], [20] studied methods to reduce mutual coherence in sparse system identification, proposing coordinate-transformation and input-design approaches that improve estimation accuracy and model order selection.

However, in many real-world applications, including control systems [21], financial time series [22], federated learning [23], and wireless communication channels [24], these assumptions are often violated. In particular, closed-loop feedback control systems seldom satisfy i.i.d. or stationarity conditions, as control inputs are inherently correlated with past outputs and disturbance processes. This discrepancy highlights the importance of developing sparse identification methods that can operate reliably under non-stationary, non-i.i.d., and feedback-dependent settings, while retaining rigorous theoretical guarantees [25].

Some initial progress has been made on sparse identification in the non-i.i.d. case. For example, [18] proposed an LS algorithm with weighted L_1 regularization that accommodates general observation sequences, and established its convergence; [26] extended this framework to multivariate ARMA systems with exogenous inputs. More recently, [27] introduced an L_γ regularization method ($0 < \gamma < 1$) for sparse parameter identification in stochastic systems, proving both exact support recovery (set convergence) and consistency of nonzero parameter estimates under general conditions. Despite these advances, a common limitation of L_1-type regularization methods has long been recognized: they can become unstable in high-dimensional settings, especially when predictors are highly correlated [4], [9]. In practice, high dimensionality almost inevitably leads to large sample correlations [28], making the collinearity problem unavoidable. In system identification—an inherently inverse problem—collinearity can cause numerical instability [29]. Moreover, estimating continuous-time impulse responses is ill-posed: small observation errors can result in large estimation deviations, and finite discretization further exacerbates this issue. Additionally, higher-order autoregressive with exogenous input (ARX) models can suffer from high variance in conventional algorithms [30].

To address ill-conditioning, classical works by [31] and [32] proposed regularization methods. Later, [33] introduced kernel-based regularization methods in reproducing kernel Hilbert spaces, with further progress in kernel design [34] and in hyperparameter estimation [35]. These approaches enhance stability mainly through L_2-type regularization, but they are not primarily designed to exploit sparsity, and thus often keep many predictors rather than yielding parsimonious models.

Motivated by these challenges, we study sparse identification of stochastic systems with non-stationary, non-i.i.d. observations and strong collinearity. We present a weighted $L_1 - L_2$ regularization scheme that performs reliable variable selection, mitigates collinearity-induced ill-conditioning, and admits rigorous guarantees. The main contributions are:

- This paper proposes a regularized two-stage adaptive identification scheme that combines weighted L_1 regularization with a quadratic penalty. The adaptive weighted L_1 term promotes sparsity and achieves exact support recovery, while the quadratic term regularizes collinearity and stabilizes high-condition-number designs. Compared with pure L_1 or quadratic-only approaches, the combined design yields more reliable variable selection under strong correlation and more stable parameter estimates.
- Leveraging martingale techniques, this paper establishes global convergence guarantees under mild excitation: (i) all zero coefficients are identified after finitely many observations with probability one; (ii) the estimates of the nonzero coefficients converge almost surely to their true values; and (iii) the estimators are asymptotically normal. Compared with existing theory, these results dispense with i.i.d. assumptions, rely on excitation conditions strictly weaker than classical persistent excitation (PE), and remain valid for closed-loop control systems.
- The group-selection property is established: when relevant regressors are highly correlated, the method retains them jointly, whereas the approach in [18] may select only one. The proposed algorithm is also applied to the identification of closed-loop linear stochastic systems with adaptive regulating control and achieves finite-time set convergence almost surely. Simulations show lower prediction mean squared error (MSE) than pure L_1 methods and yield more parsimonious models than kernel-based methods.

The remainder of the paper is organized as follows. Notation appears at the end of Section I. Section II introduces the problem formulation and regularized sparse algorithm. Section III develops the theoretical results: parameter and set convergence, asymptotic normality, and the group effect. Section IV applies the algorithm to sparse identification of linear stochastic systems with adaptive regulating control. Section V presents simulations and real examples, and Section VI concludes with a summary and future research directions.

Notation: Let $(\Omega, \mathcal{F}, \mathrm{P})$ be the probability space, $\omega \in \Omega$ be a sample point, and $E(\cdot)$ be the expectation operator. $\|\cdot\|_F$ and $\|\cdot\|$ denote Frobenius 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. I_p denotes the identity matrix of order p, $1_p = [1, ..., 1]^\top \in \mathbb{R}^p$ and $0_p = [0, ..., 0]^\top \in \mathbb{R}^p$. Moreover, sign$(\cdot)$ is defined as sign$(x) = 1$, when $x \geq 0$, and sign$(x) = -1$, when $x < 0$, and $\mathrm{vec}(x(j))|_{j=1}^{q}$ means $[x(1), x(2), \ldots, x(q)]^\top$. For the vector x, we denote its jth element by $x(j)$. For any two positive sequences $\{a_k\}$ and $\{b_k\}$, $a_k = O(b_k)$ means there exist $c > 0$ and $k_0 \in \mathbb{N}_+$ such that $a_k \leq cb_k$ for all $k \geq k_0$; $a_k = o(b_k)$ means $a_k/b_k \to 0$ as $k \to \infty$; and $a_k \asymp b_k$ means there exist $c_1, c_2 > 0$ and $k_0 \in \mathbb{N}_+$ such that $c_1 b_k \leq a_k \leq c_2 b_k$ for all $k \geq k_0$. For a symmetric matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote A's maximal and minimal eigenvalues, respectively. For two random sequences $\{x_k\}$ and $\{y_k\}$, $x_k = O_p(y_k)$ means that for any $\epsilon > 0$, there is a finite $M > 0$ and a finite $N > 0$ such that $\mathrm{P}\{|x_k| \geq M|y_k|\} < \epsilon$ for all $k \geq N$; $x_k = o_p(y_k)$ means $x_k/y_k \xrightarrow{P} 0$ as $k \to \infty$, where $\xrightarrow{P}$ means convergence in probability.

II. Problem formulation and algorithm design

A. Problem formulation

Consider the following stochastic sparse system:

$$y_{k+1} = \theta^\top \varphi_k + w_{k+1}, \quad k \geq 0, \tag{1}$$

where $\theta = [\theta(1), \ldots, \theta(p)]^\top \in \mathbb{R}^p$ is the unknown sparse parameter vector, $\varphi_k \in \mathbb{R}^p$, 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 set of zero elements of the unknown parameter θ by $A^* = \{j : \theta(j) = 0,\ j \in \{1, \ldots, p\}\}$. Suppose that there are q nonzero elements in the vector θ. Without loss of generality, we assume that $\theta(j) \neq 0$ for $j = 1, \ldots, q$ and $\theta(j) = 0$ for $j = q+1, \ldots, p$.

To proceed, we outline the assumptions used in the theoretical analysis. These conditions allow the regressor sequence φ_k to be non-i.i.d. and have broader applicability, encompassing the classical persistent excitation condition as a special case.

Assumptions. Denote the family of the σ-algebras $\mathcal{F}_k = \sigma\{y_k, \ldots, y_1, u_k, \ldots, u_0, w_k, \ldots, w_1, w'_k, \ldots, w'_1\},\ k \geq 1$, where $\{u_k\}$ and $\{w'_k\}$ are system inputs and a possible sequence of exogenous input signals, respectively. Denote the maximum and minimum eigenvalues of $\sum_{k=1}^{n} \varphi_k \varphi_k^\top$ by $\lambda_{\max}(n)$ and $\lambda_{\min}(n)$, respectively. We first give assumptions about the system noise and observation sequence.

(A1) The noise sequence $\{w_k, \mathcal{F}_k\}_{k\geq 1}$ 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) For all $k \geq 1$, φ_k is $\mathcal{F}_k$-measurable.

(A3) For the maximal and minimal eigenvalues of $\sum_{k=1}^{n} \varphi_k \varphi_k^\top$, it holds

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

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

Remark 1: Assumptions (A1) and (A2) allow the regressor sequence φ_k to be non-stationary and non-i.i.d. In (A1), the noise process $\{w_k, \mathcal{F}_k\}$ is a martingale difference sequence, which is more general than a sequence of independent random variables and imposes a weaker restriction on temporal dependence. This permits w_{k+1} to depend on $\mathcal{F}_k$ and is satisfied by many common distributions, such as Gaussian and uniform. Assumption (A2) requires that φ_k be adapted to $\{\mathcal{F}_k\}$, a condition met by a wide range of systems, including PID control, adaptive regulation, and model reference control. Assumption (A3) concerns the growth rates of the maximal and minimal eigenvalues of the sample covariance matrix of φ_k. Condition (A3)(a) corresponds to the classical weakest strong convergence requirement for LS [36], ensuring parameter convergence. Condition (A3)(b) is used for establishing set convergence in sparse identification; it is, to the best of our knowledge, the weakest known condition for this purpose and includes the traditional persistent excitation condition as a special case [18].

B. Research motivation

In high-dimensional system identification, a major challenge arises when the regressors exhibit strong correlations, which can lead to numerical instability and unreliable parameter estimates. This phenomenon, referred to as *collinearity*, is well documented in both statistical modeling [37] and large-scale system identification [38]. Formally, collinearity describes the presence of exact or near-linear dependence among a set of regressors. In the exact case, the data vectors representing k regressors lie in a subspace of dimension strictly less than k, implying that at least one regressor is an exact linear combination of the others. In practice, exact collinearity is rare, whereas near collinearity, where regressors lie approximately in such a subspace, is common and often leads to significant deterioration in estimation accuracy. A standard quantitative measure of the collinearity is the condition number $\kappa(n) = \frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}$, where $\lambda_{\max}(n)$ and $\lambda_{\min}(n)$ denote the largest and smallest eigenvalues of $\sum_{k=1}^{n} \varphi_k \varphi_k^\top$, respectively. Large values of $\kappa(n)$ indicate severe collinearity [39].

Remark 2: This measure directly relates to Assumption (A3), as the relative growth of $\lambda_{\max}(n)$ and $\lambda_{\min}(n)$ governs the estimator's convergence properties. In particular, these conditions allow for certain moderate collinearity scenarios under which the sample condition number $\kappa(n) = \lambda_{\max}(n)/\lambda_{\min}(n)$ may still diverge. For example, if $\lambda_{\min}(n) \asymp n$ and $\lambda_{\max}(n) \asymp n \log n$, then $\kappa(n) \asymp \log n \to \infty$ while the assumptions remain satisfied.

In words, collinearity means that some regressors are almost linear combinations of others, so the matrix $\sum_{k=1}^{n} \varphi_k \varphi_k^\top$ will have a large condition number. This leads to inflated variance in the parameter estimates, instability in the solution path, and difficulty in correctly selecting relevant variables.

To address these challenges, most existing approaches build upon the classical LS estimation framework, incorporating suitable regularization to improve stability and interpretability. Given the observed data $\{y_{k+1}, \varphi_k\}_{k=1}^{n}$, the LS estimate of the parameter vector $\theta_{n+1} = [\theta_{n+1}(1), \ldots, \theta_{n+1}(p)]^\top$ is

$$\theta_{n+1} = \left(\sum_{k=1}^{n} \varphi_k \varphi_k^\top\right)^{-1} \left(\sum_{k=1}^{n} \varphi_k y_{k+1}\right). \tag{2}$$

Regularized LS with weighted L_1 penalties. For sparse parameter identification, [18] proposed a weighted L_1 regularization method based on the LS estimate, solving

$$J_n^1(\beta) = \sum_{k=1}^{n} \left(y_{k+1} - \beta^\top \varphi_k\right)^2 + \lambda_n \sum_{j=1}^{p} \frac{1}{|\hat{\theta}_{n+1}(j)|} |\beta(j)|, \tag{3}$$

where $\lambda_n > 0$ is a tuning parameter, and

$$\hat{\theta}_{n+1}(j) = \theta_{n+1}(j) + \text{sign}(\theta_{n+1}(j)) \sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}},\ j = 1, \ldots, p.$$

While this approach induces sparsity and enjoys favorable asymptotic properties under non-i.i.d. and non-stationary observations, relying solely on an L_1 penalty can lead to instability in the presence of collinearity [9], and it lacks the *grouping effect*, often selecting only a single variable from a set of highly correlated regressors.

Regularized LS with quadratic penalties. To improve numerical stability in ill-conditioned problems, [29], [30] introduced a quadratic regularization term $\beta^\top P \beta$ with $P > 0$:

$$J_n^2(\beta) = \sum_{k=1}^{n} \left(y_{k+1} - \beta^\top \varphi_k\right)^2 + \bar{\lambda}_n \beta^\top P \beta, \tag{4}$$

where $\bar{\lambda}_n > 0$ controls the trade-off between data fitting and the quadratic penalty. This quadratic term effectively mitigates variance inflation in the estimates and enhances the stability

of the estimates under collinearity [29], but it does not directly promote sparsity.

Identification objective. Motivated by these observations, our aim is to develop a unified estimation framework that simultaneously achieves variable selection, accurate parameter estimation, and improved predictive performance for stochastic sparse systems under non-i.i.d., non-stationary, and ill-conditioned settings. The proposed methodology seeks to consistently identify the zero coefficients A^* and estimate the nonzero coefficients of the unknown parameter vector from $\{y_{k+1}, \varphi_k\}_{k=1}^n$, while providing theoretical guarantees of parameter convergence, set consistency, and asymptotic normality.

C. Regularized sparse identification algorithm

In this subsection, we develop a *two-stage adaptively weighted regularized* method that integrates the advantages of sparsity-promoting penalties with stability-enhancing quadratic regularization, while being tailored for non-i.i.d., non-stationary, and potentially ill-conditioned settings.

Stage 1 (stable preliminary estimation). Motivated by the numerical stabilization effects of kernel-based quadratic regularization [29], [30], we first solve the following penalized LS problem:

$$J_n^0(\beta) = \sum_{k=1}^{n} \left(y_{k+1} - \beta^\top \varphi_k\right)^2 + \lambda_{1n}^{(1)} \sum_{j=1}^{p} |\beta(j)| + \lambda_{2n} \beta^\top P_n \beta, \tag{5}$$

where $\lambda_{1n}^{(1)} > 0$ enforces sparsity, $\lambda_{2n} > 0$ controls the quadratic shrinkage, and $P_n > 0$ is a design matrix reflecting prior structural information. The initial estimate is obtained as

$$\beta_n^0 = \arg\min_{\beta \in \mathbb{R}^p} J_n^0(\beta). \tag{6}$$

Stage 2 (adaptive sparsity refinement). Since a plain L_1 penalty requires strong irrepresentable conditions for consistent variable selection [40], we adaptively reweight the L_1 term using the Stage-1 estimate (6), in the spirit of [18]. To prevent degeneracy when $\beta_n^0(j)$ is close to zero, we introduce a stabilizing offset that depends on the spectral properties of $\sum_{k=1}^n \varphi_k \varphi_k^\top$ and P_n:

$$\begin{aligned} \hat{\beta}_n^0(j) = &\beta_n^0(j) + \operatorname{sign}(\beta_n^0(j)) \left(\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} \right. \\ &\left. + \frac{\lambda_{1n}^{(1)} + \lambda_{2n} \lambda_{\max}(P_n)}{\lambda_{\min}\left(\sum_{k=1}^n \varphi_k \varphi_k^\top + \lambda_{2n} P_n\right)} \right). \end{aligned} \tag{7}$$

The refined problem is then

$$\begin{aligned} \hat{J}_n(\beta) = &\sum_{k=1}^{n} \left(y_{k+1} - \beta^\top \varphi_k\right)^2 + \lambda_{1n}^{(2)} \sum_{j=1}^{p} \frac{1}{|\hat{\beta}_n^0(j)|} |\beta(j)| \\ &+ \lambda_{2n} \beta^\top P_n \beta, \end{aligned} \tag{8}$$

yielding the final estimator

$$\hat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^p} \hat{J}_n(\beta), \ \widehat{A}_n^* = \{j \in \{1, \ldots, p\} \,|\, \hat{\beta}_n(j) = 0\}. \tag{9}$$

This two-stage design allows the first stage to stabilize the estimation under collinearity, while the second stage sharpens sparsity recovery with adaptive weighting, preserving both numerical robustness and variable selection consistency.

Conditions on regularization parameters. To establish convergence, sparsity recovery, and asymptotic normality, we impose the following conditions on the regularization sequences.

(A4) The positive regularization parameters $\{\lambda_{1n}^{(1)}\}$, $\{\lambda_{1n}^{(2)}\}$ and $\{\lambda_{2n}\}$ in (5) and (8) satisfy the following:

(a) $\dfrac{\lambda_{1n}^{(1)} + \lambda_{1n}^{(2)} + \lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n) + \lambda_{2n}\lambda_{\min}(P_n)} \xrightarrow[n\to\infty]{} 0,$

(b) $\lambda_{2n}\lambda_{\max}(P_n) = O(\lambda_{1n}^{(2)}),$

(c) $\dfrac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)} \cdot \dfrac{\lambda_{\max}(n)}{\lambda_{1n}^{(2)}} \xrightarrow[n\to\infty]{} 0,$

(d) $\dfrac{\lambda_{1n}^{(1)} + \lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}} \cdot \dfrac{\lambda_{\max}(n)}{\lambda_{\min}(n)} \cdot \sqrt{\dfrac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} \xrightarrow[n\to\infty]{} 0,$

(e) $\dfrac{\left(\lambda_{1n}^{(1)} + \lambda_{2n}\lambda_{\max}(P_n)\right)\lambda_{\max}(n)}{\lambda_{\min}(n)^2} \xrightarrow[n\to\infty]{} 0.$

These conditions balance the sparsity-inducing and stability-enhancing effects of the penalties, ensuring well-posedness of the intermediate problems and enabling the theoretical guarantees established in Section III.

Remark 3: Assumption (A4)(a) ensures parameter convergence of the proposed algorithm, while Assumption (A4)(b) allows adaptation to a broader range of scenarios. The overall Assumption (A4) is imposed to guarantee set convergence. Under the modified formulation in which λ_{2n} appears in the form $\lambda_{2n}\lambda_{\max}(P_n)$ and $\lambda_{\min}(n) + \lambda_{2n}\lambda_{\min}(P_n)$ in the denominators, Assumption (A4) can be fulfilled with the following simple choice:

$$\lambda_{2n}\lambda_{\max}(P_n) = \sqrt{\lambda_{\min}(n)}, \quad \lambda_{1n}^{(1)} = \lambda_{\min}(n)\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}},$$
$$\lambda_{1n}^{(2)} = \lambda_{\max}(n)\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}}. \tag{10}$$

Indeed, noting that $\frac{\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(1)}} = \frac{1}{\sqrt{\log \lambda_{\max}(n)}} = o(1)$, $\frac{\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}} = \frac{\lambda_{\min}(n)}{\lambda_{\max}(n)} \frac{1}{\sqrt{\log \lambda_{\max}(n)}} = o(1)$, a direct calculation shows that $\frac{\lambda_{1n}^{(1)} + \lambda_{1n}^{(2)} + \lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n) + \lambda_{2n}\lambda_{\min}(P_n)} = O\left(\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}}\right) = o(1)$, $\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)} \cdot \frac{\lambda_{\max}(n)}{\lambda_{1n}^{(2)}} = O\left(\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}}\right) = o(1)$, $\frac{\lambda_{1n}^{(1)} + \lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}} \cdot \frac{\lambda_{\max}(n)}{\lambda_{\min}(n)} \cdot \sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} = O\left(\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}}\right) = o(1)$, $\frac{\left(\lambda_{1n}^{(1)} + \lambda_{2n}\lambda_{\max}(P_n)\right)\lambda_{\max}(n)}{\lambda_{\min}(n)^2} = O\left(\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}}\right) = o(1)$.

The specific algorithm is shown in Algorithm 1.

Remark 4: The roles of $\lambda_{1n}^{(1)}$, $\lambda_{1n}^{(2)}$, and λ_{2n} can be distinguished as follows. The parameters $\lambda_{1n}^{(1)}$ and $\lambda_{1n}^{(2)}$ in (11) and (13) promote sparsity in the estimated coefficients, whereas λ_{2n} is designed to improve estimation accuracy under collinearity or other ill-conditioned settings. When both λ_{2n} and $\lambda_{1n}^{(1)}$ are set to zero in (11) and (13), the procedure in (14) reduces to the algorithm of [18].

Remark 5: We introduce a general quadratic regularization matrix P_n (rather than I_p) because a structured P_n can encode

Algorithm 1 Regularized sparse Algorithm

Step 0 (Initialization). Choose positive sequences $\{\lambda_{1n}^{(1)}\}_{n\geq 1}$, $\{\lambda_{1n}^{(2)}\}_{n\geq 1}$ and $\{\lambda_{2n}\}_{n\geq 1}$ satisfying Assumption (A4) or simply choose as (10).

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

$$\beta_n^0 = \underset{\beta}{\operatorname{argmin}}\{\sum_{k=1}^{n}\left(y_{k+1}-\beta^\top\varphi_k\right)^2+\lambda_{1n}^{(1)}\sum_{j=1}^{p}|\beta(j)| + \lambda_{2n}\beta^\top P_n\beta\}. \tag{11}$$

Let $\beta_n^0=\left[\beta_n^0(1),\ldots,\beta_n^0(p)\right]^\top$, and for $1\leq j\leq p$, define

$$\hat{\beta}_n^0(j)=\beta_n^0(j)+\operatorname{sign}(\beta_n^0(j))\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}} + \frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}\left(\sum_{k=1}^n\varphi_k\varphi_k^\top+\lambda_{2n}P_n\right)}\right). \tag{12}$$

Step 2 (Weighted regularization estimate). With $\lambda_{1n}^{(1)}$, $\lambda_{1n}^{(2)}$, λ_{2n}, and $\hat{\beta}_n^0(j)$, optimize the objective function

$$\hat{J}_n(\beta)=\sum_{k=1}^{n}\left(y_{k+1}-\beta^\top\varphi_k\right)^2+\lambda_{1n}^{(2)}\sum_{j=1}^{p}\frac{1}{|\hat{\beta}_n^0(j)|}|\beta(j)| + \lambda_{2n}\beta^\top P_n\beta \tag{13}$$

and obtain

$$\begin{cases}\widehat{\beta}_n=\left[\widehat{\beta}_n(1),\ldots,\widehat{\beta}_n(p)\right]^\top=\underset{\beta}{\operatorname{argmin}}\ \hat{J}_n(\beta),\\ \widehat{A}_n^*=\left\{j=1,\ldots,p\mid\widehat{\beta}_n(j)=0\right\}.\end{cases} \tag{14}$$

prior information, often improving conditioning and finite-sample stability while preserving convexity; in this sense, P_n is more expressive and task-aware than I_p. Typical choices include first-order difference [41], Laplacian-based regularization [42], and covariance–shrinkage forms [43]. Statistically, such P_n helps control variance and improves the bias–variance trade-off under collinearity [44]; from an optimization viewpoint, any $P_n>0$ maintains strict convexity and induces a better-conditioned proximal geometry [45]; numerically, it improves the conditioning of the regularized information sample matrix and stabilizes estimation. These choices do not alter the asymptotic properties established in Section III-A.

Remark 6: The quadratic regularization term $\lambda_{2n}\beta^\top P_n\beta$ in (11) and (13) may may introduce a small finite-sample shrinkage effect, making $\widehat{\beta}_n$ asymptotically unbiased rather than exactly unbiased. A simple bias-reduction adjustment can be applied by defining $\widehat{\widehat{\beta}}_n = \left(I_p+\lambda_{2n}\left(\sum_{k=1}^n\varphi_k\varphi_k^\top\right)^{-1}P_n\right)\widehat{\beta}_n$. When $P_n=I_p$, this can be reduced to $\widehat{\widehat{\beta}}_n=\left(1+\frac{\lambda_{2n}}{\lambda_{\min}(n)}\right)\widehat{\beta}_n$. This adjustment is consistent with the results in Subsection III-A and preserves the asymptotic properties of the original estimator.

Remark 7: The set $\widehat{A}_n^*$ in (14) is intended to identify the indices corresponding to zero coefficients in the parameter vector. In numerical computation, however, the exact minimizer of (13) may not yield coefficients that are numerically *exactly* zero due to finite precision and optimization tolerances. To address this, one can introduce a small threshold $\epsilon>0$ (e.g., $\epsilon=10^{-10}$) and redefine $\widehat{A}_n^*=\left\{j\in\{1,\ldots,p\}\,\middle|\,|\widehat{\beta}_n(j)|<\epsilon\right\}$. This practical adjustment ensures stable and reliable variable selection.

III. Main Results of Regularized Sparse Algorithm

This section establishes the theoretical properties of Algorithm 1 for the non-stationary and non-i.i.d. setting. Subsection III-A presents asymptotic convergence results, including parameter consistency and exact identification of zero coefficients with finitely many observations. Subsection III-B derives the asymptotic normality, and Subsection III-C investigates the group effect. While the beneficial effects of incorporating a quadratic regularization matrix P_n have been well recognized in the literature (see, e.g., [29]), and various approaches for selecting P_n based on prior information have been explored in related works, the focus here is on rigorously establishing large-sample convergence properties for general choices of P_n. This generality allows the analysis to cover a broad range of practical choices for P_n, and extends the results beyond the traditional stationary and i.i.d. assumptions to the non-stationary and non-i.i.d. settings.

A. Asymptotic convergence

For the estimate $\hat{\beta}_n$ and $\widehat{A}_n^*$ generated by Algorithm 1, we have the following results:

Theorem 1 (Parameter convergence): If Assumptions (A1), (A2), (A3)(a) and (A4)(a) hold, the estimate $\hat{\beta}_n$ generated by Algorithm 1 converges almost surely, i.e., $\hat{\beta}_n(i)\xrightarrow[n\to\infty]{}\theta(i)$ for $i=1,\ldots,p$ a.s.

Theorem 2 (Set convergence): Let $\hat{\beta}_n=(\hat{\beta}_{1n}^\top,\hat{\beta}_{2n}^\top)^\top$ with $\hat{\beta}_{1n}\in\mathbb{R}^q$ and $\hat{\beta}_{2n}\in\mathbb{R}^{p-q}$ denoting, respectively, the subvectors formed by the first q and the other components of $\hat{\beta}_n$. Let $\theta=(\theta_{10}^\top,0_{p-q}^\top)^\top$ with $\theta_{10}\in\mathbb{R}^q$. If Assumptions (A1), (A2), (A3)(b), and (A4) hold, then there exists a set Ω_0 with $\mathbb{P}(\Omega_0)=1$ such that, for every $\omega\in\Omega_0$, there exists an integer $N_0(\omega)$ such that $\widehat{A}_n^*=A^*,\ \forall\ n\geq N_0(\omega)$.

Theorem 1 establishes that the parameter estimates converge to the true values, while Theorem 2 shows that the index set of zero components in θ can be identified with probability one after finitely many observations. We now proceed to prove Theorems 1 and 2. To this end, we recall two classical results that will be used in the analysis.

Lemma 1: [46] Consider system (1). If Assumptions (A1) and (A2) hold, then

$$\left\|\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\right)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k w_{k+1}\right\|=O\left(\sqrt{\log\lambda_{\max}(n)}\right),\quad\text{a.s.}$$

Lemma 2: [46] Consider system (1). If Assumptions (A1) and (A2) hold, then the estimation error of the LS algorithm

in (2) satisfies

$$\|\theta_{n+1}-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\right), \quad \text{a.s.}$$

Since Algorithm 1 involves two sequential steps, the convergence properties are not straightforward to analyze. To facilitate the analysis, we first consider the minimizer of a generalized objective function associated with (8) and derive a technical lemma on the non-asymptotic behavior of the corresponding estimate. Given observations $\{y_{k+1},\varphi_k\}_{k=1}^n$, regularization parameters $\alpha_{1n},\ \alpha_{2n}>0$, regularization matrix $P_n>0$, and a weight vector $\eta_n=[\eta_n(1),\ldots,\eta_n(p)]^\top\in\mathbb{R}^p$ with strictly positive components, we define the following auxiliary objective function and its minimizer:

$$J_n(\beta,\alpha_{1n},\alpha_{2n},\eta_n)=\sum_{k=1}^n\left(y_{k+1}-\beta^\top\varphi_k\right)^2+\alpha_{2n}\beta^\top P_n\beta + \alpha_{1n}\sum_{j=1}^p\eta_n(j)|\beta(j)|, \tag{15}$$

$$\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)=\operatorname*{argmin}_\beta J_n(\beta,\alpha_{1n},\alpha_{2n},\eta_n). \tag{16}$$

Lemma 3: Under Assumptions (A1) and (A2), the auxiliary estimate $\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)$ defined in (16) admits the following non-asymptotic bound:

$$\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\alpha_{1n}\sqrt{\sum_{j=1}^q\eta_n^2(j)}+\alpha_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\right). \tag{17}$$

Proof: The proof is given in Appendix. ∎

From Lemma 3, we directly obtain a non-asymptotic error bound of the estimate in (12) which yields parameter convergence for the Stage 1 estimate and contributes to the overall convergence of Algorithm 1.

Corollary 1: Under Assumptions (A1) and (A2), the initial estimate $\hat\beta_n^0$ defined in (12) satisfies

$$\|\hat\beta_n^0-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\right). \tag{18}$$

Proof: From (5), (6), (15), and (16), we have $\beta_n^0=\beta_n(\lambda_{1n}^{(1)},\lambda_{2n},1_p)$. The bound (18) follows directly from Lemma 3 together with (12). ∎

We now present the proofs of Theorems 1 and 2.

Proof of Theorem 1. From (8) and (16), define $\hat\eta_n(i):=1/|\hat\beta_n^0(i)|$, $i=1,\ldots,p$, and $\hat\eta_n:=(\hat\eta_n(1),\ldots,\hat\eta_n(p))^\top$. We then have $\hat\beta_n=\beta_n\big(\lambda_{1n}^{(2)},\,\lambda_{2n},\,\hat\eta_n\big)$. Moreover, by Assumptions (A3)(a), (A4)(a), and Corollary 1, we know that $\beta_n^0(i)\to\theta(i)$ and $\hat\beta_n^0(i)\to\theta(i)$ for $i=1,\ldots,q$, where each $\theta(i)\neq 0$. Thus, for sufficiently large n, $\beta_n^0(i)$ and $\hat\beta_n^0(i)$ are both bounded away from zero. Hence $\sum_{j=1}^q\hat\eta_n(j)^2=O(1)$, which fits the bound in Lemma 3.

$$\|\hat\beta_n-\theta\| = \|\beta_n(\lambda_{1n}^{(2)},\,\lambda_{2n},\,\hat\eta_n)-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\right).$$

Therefore, by Assumptions (A3)(a) and (A4)(a), this completes the proof. ∎

Proof of Theorem 2. Let $t_n:=\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}+\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}$. Decompose the estimate $\hat\beta_n=(\hat\beta_{1n}^\top,\,\hat\beta_{2n}^\top)^\top$ as $\hat\beta_{1n}=\theta_{10}+t_n\,u_{1n},\quad \hat\beta_{2n}=t_n\,u_{2n}$, where $u_{1n}\in\mathbb{R}^q$ and $u_{2n}\in\mathbb{R}^{p-q}$. In addition, denote

$$\sum_{k=1}^n\varphi_k\varphi_k^\top=\begin{bmatrix}\Phi_n^{(11)} & \Phi_n^{(12)}\\ \Phi_n^{(21)} & \Phi_n^{(22)}\end{bmatrix},\quad \varphi_k=\begin{bmatrix}\varphi_k^{(1)}\\ \varphi_k^{(2)}\end{bmatrix}, \tag{19}$$

where $\Phi_n^{(11)}\in\mathbb{R}^{q\times q}$, $\varphi_k^{(1)}\in\mathbb{R}^q$, and other blocks have compatible dimensions. By Lemma 3, there exists $\bar U>0$ such that $\|\hat\beta_n-\theta\|\le t_n\bar U$. Thus, $\|u_{1n}\|\le\bar U$. Since $\hat\beta_n$ is the minimizer of $\hat J_n(\beta)$ in (13), it follows that

$$\hat J_n(\hat\beta_n)-\hat J_n(\hat\beta_{1n},\,0_{p-q})\le 0. \tag{20}$$

Noting (1), a direct calculation for (13) yields

$$\begin{aligned}\hat J_n(\hat\beta_n) &= \sum_{k=1}^n w_{k+1}^2+(\hat\beta_n-\theta)^\top\left(\sum_{k=1}^n\varphi_k\varphi_k^\top\right)(\hat\beta_n-\theta)\\ &\quad+2\sum_{k=1}^n\varphi_k^\top(\theta-\hat\beta_n)w_{k+1}+\lambda_{2n}\hat\beta_n^\top P_n\hat\beta_n\\ &\quad+\lambda_{1n}^{(2)}\sum_{j=1}^p\frac{1}{|\hat\beta_n^0(j)|}\left|\hat\beta_n(j)\right|\\ &=\sum_{k=1}^n w_{k+1}^2+t_n^2u_{1n}^\top\Phi_n^{(11)}u_{1n}+t_n^2u_{2n}^\top\Phi_n^{(22)}u_{2n}\\ &\quad+2t_n^2\sum_{k=1}^n\big(\varphi_k^{(1)\top}u_{1n}\big)\big(\varphi_k^{(2)\top}u_{2n}\big)\\ &\quad-2t_n\sum_{k=1}^n\varphi_k^{(1)\top}u_{1n}w_{k+1}-2t_n\sum_{k=1}^n\varphi_k^{(2)\top}u_{2n}w_{k+1}\\ &\quad+\lambda_{2n}t_n^2\big(u_{1n}^\top P_{n,11}u_{1n}+2u_{1n}^\top P_{n,12}u_{2n}+u_{2n}^\top P_{n,22}u_{2n}\big)\\ &\quad+\lambda_{1n}^{(2)}\sum_{j=1}^q\frac{1}{|\hat\beta_n^0(j)|}\left|\hat\beta_{1n}(j)\right|+\lambda_{1n}^{(2)}t_n\sum_{j=1}^{p-q}\frac{|u_{2n}(j)|}{|\hat\beta_n^0(j+q)|},\end{aligned} \tag{21}$$

where P_n is block-partitioned as $P_n=\begin{bmatrix}P_{n,11} & P_{n,12}\\ P_{n,21} & P_{n,22}\end{bmatrix}$ with conformable dimensions. Similarly,

$$\begin{aligned}&\hat J_n(\hat\beta_{1n},0_{p-q})\\ &=\sum_{k=1}^n w_{k+1}^2+t_n^2u_{1n}^\top\Phi_n^{(11)}u_{1n}-2t_n\sum_{k=1}^n\varphi_k^{(1)\top}u_{1n}w_{k+1}\\ &\quad+\lambda_{2n}t_n^2u_{1n}^\top P_{n,11}u_{1n}+\lambda_{1n}^{(2)}\sum_{j=1}^q\frac{1}{|\hat\beta_n^0(j)|}\left|\hat\beta_{1n}(j)\right|.\end{aligned} \tag{22}$$

Subtracting (22) from (21), we obtain

$$\begin{aligned}\hat J_n(\hat\beta_n)-\hat J_n(\hat\beta_{1n},0_{p-q}) &= 2t_n^2\sum_{k=1}^n\big(\varphi_k^{(1)\top}u_{1n}\big)\big(\varphi_k^{(2)\top}u_{2n}\big)\\ &\quad+t_n^2u_{2n}^\top\Phi_n^{(22)}u_{2n}-2t_n\sum_{k=1}^n\varphi_k^{(2)\top}u_{2n}w_{k+1}\\ &\quad+\lambda_{2n}t_n^2\big(2u_{1n}^\top P_{n,12}u_{2n}+u_{2n}^\top P_{n,22}u_{2n}\big)\\ &\quad+\lambda_{1n}^{(2)}t_n\sum_{j=1}^{p-q}\frac{|u_{2n}(j)|}{|\hat\beta_n^0(j+q)|}\\ &\triangleq T_n^{(1)}+T_n^{(2)}+T_n^{(3)}+T_n^{(4)}+T_n^{(5)}.\end{aligned} \tag{23}$$

We then separately estimate $T_n^{(i)}$, $i=1,\dots,5$. First, using $\|u_{1n}\| \le \bar{U}$ and $\|A\| \le \|A\|_F \le c_2\|A\|$ for some constant $c_2>0$, we have

$$\begin{aligned}
&T_n^{(1)}+T_n^{(2)}\\
&=2t_n^2\sum_{k=1}^{n}\big(\varphi_k^{(1)\top}u_{1n}\big)\big(\varphi_k^{(2)\top}u_{2n}\big)+t_n^2\sum_{k=1}^{n}\big(\varphi_k^{(2)\top}u_{2n}\big)^2\\
&\ge 2t_n^2\,u_{1n}^\top\left(\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(2)\top}\right)u_{2n}\\
&\ge -2t_n^2\|u_{1n}\|\,\|u_{2n}\|\left\|\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(2)\top}\right\|\\
&\ge -2t_n^2\bar{U}\,\|u_{2n}\|\left\|\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(2)\top}\right\|_F\\
&\ge -2t_n^2\bar{U}\,\|u_{2n}\|\left\|\sum_{k=1}^{n}\varphi_k\varphi_k^{\top}\right\|_F\\
&\ge -2t_n^2\bar{U}\,\|u_{2n}\|\,c_2\left\|\sum_{k=1}^{n}\varphi_k\varphi_k^{\top}\right\|\\
&\ge -2c_2\,t_n^2\bar{U}\,\|u_{2n}\|\,\lambda_{\max}(n)
\end{aligned}\tag{24}$$

Second, by Lemma 1 and the bound $\lambda_{\max}\{\Phi_n^{(22)}\} \le \lambda_{\max}(n)$, we have

$$\begin{aligned}
T_n^{(3)}&=-2t_n\sum_{k=1}^{n}\varphi_k^{(2)^\top}\,u_{2n}\,w_{k+1}\\
&=-2t_n\,u_{2n}^\top\Big(\Phi_n^{(22)}\Big)^{\frac12}\Big(\Phi_n^{(22)}\Big)^{-\frac12}\sum_{k=1}^{n}\varphi_k^{(2)}w_{k+1}\\
&\ge -2t_n\|u_{2n}\|\big\|\Phi_n^{(22)\,\frac12}\big\|\left\|\Big(\Phi_n^{(22)}\Big)^{-\frac12}\sum_{k=1}^{n}\varphi_k^{(2)}w_{k+1}\right\|\\
&\ge -2t_n\|u_{2n}\|\,\lambda_{\max}(n)^{\frac12}\log^{\frac12}\lambda_{\max}(n).
\end{aligned}\tag{25}$$

Third, define $\gamma_n := \sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}$. By the definition of $\hat{\beta}_n^0$, Corollary 1, and the fact that $\theta(j)=0$ for $j=q+1,\dots,p$, there exist constants $c_4>c_3\ge 1$ such that $c_3\gamma_n \le |\hat{\beta}_n^0(j)| \le c_4\gamma_n,\ j=q+1,\dots,p$, which implies

$$\frac{1}{c_4\gamma_n}\le\frac{1}{|\hat{\beta}_n^0(j)|}\le\frac{1}{c_3\gamma_n},\quad j=q+1,\dots,p.\tag{26}$$

Therefore, using (26) and the block partition of P_n, we obtain

$$\begin{aligned}
&T_n^{(4)}+T_n^{(5)}\\
&=\lambda_{2n}t_n^2\big(2u_{1n}^\top P_{n,12}u_{2n}+u_{2n}^\top P_{n,22}u_{2n}\big)+\lambda_{1n}^{(2)}t_n\sum_{j=1}^{p-q}\frac{|u_{2n}(j)|}{|\hat{\beta}_n^0(j+q)|}\\
&\ge -2\lambda_{2n}t_n^2\bar{U}\,\|P_{n,12}\|\,\|u_{2n}\|+\frac{\lambda_{1n}^{(2)}t_n}{c_4\gamma_n}\,\|u_{2n}\|.
\end{aligned}$$

Now we prove that $\|P_{n,12}\| \le \lambda_{\max}(P_n)$. By the Cauchy-Schwarz inequality, for arbitrary vectors $x\in\mathbb{R}^q$ and $y\in\mathbb{R}^{p-q}$, we have $|x^\top P_{n,12}y| \le \sqrt{x^\top P_{n,11}x}\cdot\sqrt{y^\top P_{n,22}y}$. Taking the supremum over x and y such that $\|x\|=\|y\|=1$, we get $\|P_{n,12}\| = \sup_{\|x\|=\|y\|=1}|x^\top P_{n,12}y| \le \sup_{\|x\|=\|y\|=1}\sqrt{x^\top P_{n,11}x}\cdot\sqrt{y^\top P_{n,22}y}$. Since $\|x\|=\|y\|=1$, this simplifies to $\|P_{n,12}\| \le \sqrt{\|P_{n,11}\|\cdot\|P_{n,22}\|}$. Next, because $\|P_{n,11}\|\le\|P_n\|$ and $\|P_{n,22}\|\le\|P_n\|$, we obtain: $\|P_{n,12}\|\le\sqrt{\|P_n\|\cdot\|P_n\|}=\|P_n\|=\lambda_{\max}(P_n)$. Substituting this result into the expression for $T_n^{(4)}+T_n^{(5)}$ yields

$$T_n^{(4)}+T_n^{(5)} \ge -2\lambda_{2n}t_n^2\bar{U}\lambda_{\max}(P_n)\|u_{2n}\|+\frac{\lambda_{1n}^{(2)}t_n}{c_4\gamma_n}\|u_{2n}\|.\tag{27}$$

Combining (23) and (27), we have the following overall lower bound for the terms:

$$\begin{aligned}
&T_n^{(1)}+T_n^{(2)}+T_n^{(3)}+T_n^{(4)}+T_n^{(5)}\\
&\ge -2c_2t_n^2\bar{U}\|u_{2n}\|\lambda_{\max}(n)-2\lambda_{2n}t_n^2\bar{U}\lambda_{\max}(P_n)\|u_{2n}\|\\
&\quad+\frac{\lambda_{1n}^{(2)}t_n}{c_4\gamma_n}\|u_{2n}\|-2t_n\|u_{2n}\|\lambda_{\max}(n)^{\frac12}\log^{\frac12}\lambda_{\max}(n)\\
&=\frac{\lambda_{1n}^{(2)}t_n}{c_4\gamma_n}\|u_{2n}\|\left(1-2c_2c_4\bar{U}\,\frac{t_n\gamma_n\,\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}-2c_4\bar{U}t_n\gamma_n\,\lambda_{2n}\right.\\
&\quad\left.\cdot\frac{\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}-2c_4\,\frac{\gamma_n\,\lambda_{\max}(n)^{\frac12}\log^{\frac12}\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\right).
\end{aligned}\tag{28}$$

Hence, by (20), (23), and (28), we have

$$\begin{aligned}
0\ge&\|u_{2n}\|\left(1-2c_2c_4\bar{U}\,\frac{t_n\gamma_n\,\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}-2c_4\bar{U}t_n\gamma_n\,\lambda_{2n}\right.\\
&\left.\cdot\frac{\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}-2c_4\,\frac{\gamma_n\,\lambda_{\max}(n)^{\frac12}\log^{\frac12}\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\right).
\end{aligned}\tag{29}$$

On the other hand, recall that

$$t_n=\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)},$$

$$\gamma_n=\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}.$$

We now bound the three terms inside the parentheses of (29) separately.

(i) First term. Expanding $t_n\gamma_n$ termwise, we obtain

$$\begin{aligned}
&\frac{t_n\gamma_n\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\le\frac{\log\lambda_{\max}(n)\lambda_{\max}(n)}{\lambda_{\min}(n)\lambda_{1n}^{(2)}}\\
&+\frac{\sqrt{\log\lambda_{\max}(n)}\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)}{\lambda_{\min}(n)^{\frac32}\lambda_{1n}^{(2)}}\\
&+\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\,\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\\
&+\frac{\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)}{\lambda_{\min}(n)^2}\,\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\\
&\xrightarrow[n\to\infty]{}0\ \text{a.s.}
\end{aligned}\tag{30}$$

Indeed, the four terms on the right tend to zero by (A4)(c), (A4)(d), (A3)(b)+ (A4)(b), and (A4)(e)+ (A4)(b), respectively.
(ii) Second term. For the mixed factor with $\lambda_{2n}\lambda_{\max}(P_n)$, for some constant $C>0$, we obtain

$$\begin{aligned}
&\frac{t_n\gamma_n\,\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\le C\left[\frac{\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\right.\\
&\left.+\frac{\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\cdot\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\right]\\
&\xrightarrow[n\to\infty]{}0\ \text{a.s. (by (A3)(a),(A4)(b) and (A4)(a),(A4)(b))}
\end{aligned}\tag{31}$$

(iii) Third term. Finally,

$$\begin{aligned}
&\frac{\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)\gamma_n}{\lambda_{1n}^{(2)}}\\
\le\ &\frac{\log\lambda_{\max}(n)\lambda_{\max}(n)^{\frac{1}{2}}}{\lambda_{\min}(n)^{\frac{1}{2}}\lambda_{1n}^{(2)}}\\
&+\frac{\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)}{\lambda_{\min}(n)\lambda_{1n}^{(2)}}\\
=\ &\frac{\log\lambda_{\max}(n)\lambda_{\max}(n)}{\lambda_{\min}(n)\lambda_{1n}^{(2)}}\frac{\lambda_{\min}(n)^{\frac{1}{2}}}{\lambda_{\max}(n)^{\frac{1}{2}}}\\
&+\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\frac{\lambda_{\min}(n)^{\frac{1}{2}}}{\lambda_{\max}(n)^{\frac{1}{2}}}\\
&\xrightarrow[n\to\infty]{} 0 \text{ a.s. (by (A4)(c) and (A4)(d))} \qquad (32)
\end{aligned}$$

Combining (30), (31), and (32), we deduce that each of the three terms inside the parentheses of (29) converges to zero a.s. Hence, for all $\omega\in\Omega_0$ with $\mathrm{P}(\Omega_0)=1$,

$$\begin{aligned}
&1-2c_2c_4\bar{U}\,\frac{t_n\gamma_n\,\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}-2c_4\bar{U}\,\frac{t_n\gamma_n\,\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\\
&-2c_4\,\frac{\gamma_n\,\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\ >\ 0.
\end{aligned}$$

Substituting this into (29) yields $\|u_{2n}\|=0$ for all sufficiently large n, a contradiction unless $\widehat{A}_n^*=A^*$. This completes the proof. ■

B. Asymptotic normality

In this section, we show that the estimates of the nonzero elements of the parameter vector generated by Algorithm 1 are asymptotically normal.

Theorem 3: Assume for each n that there exists a deterministic symmetric positive definite matrix R_n such that

$$\begin{aligned}
&R_n^{-1}\Phi_n^{(11)}\xrightarrow{P} I_q,\quad \max_{1\le k\le n}\|R_n^{-1/2}\varphi_k^{(1)}\|\xrightarrow{P}0, \qquad (33)\\
&\lim_{k\to\infty}E\left(w_{k+1}^2\mid\mathcal{F}_k\right)=\sigma^2 \text{ a.s. for some constant } \sigma,
\end{aligned}$$

where $\varphi_k^{(1)}$ and $\Phi_n^{(11)}$ are defined in (19). Write the estimate $\hat{\beta}_n=(\hat{\beta}_{1n}^\top,\hat{\beta}_{2n}^\top)^\top$ and $\theta=[\theta_{10}^\top,0_{p-q}^\top]^\top$. For any non-random $\upsilon_n\in\mathbb{R}^q$ satisfying $\|\upsilon_n\|\le 1$, let $s_n^2=\sigma^2\,\upsilon_n^\top R_n^{-1}\upsilon_n$. If Assumptions (A1)–(A4) hold and $\lambda_{1n}^{(2)}\lambda_{\min}(n)^{-\frac{1}{2}}\xrightarrow{P}0$, then

$$s_n^{-1}\,\upsilon_n^\top\left(\left(I_q+\lambda_{2n}\,(\Phi_n^{(11)})^{-1}P_{n,11}\right)\hat{\beta}_{1n}-\theta_{10}\right)\xrightarrow{\mathrm{d}}N(0,1), \qquad (34)$$

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

Remark 8: The deterministic matrix R_n, satisfying (33) in Theorem 3, can be regarded as a stability assumption for the matrix $\Phi_n^{(11)}$. This assumption plays a critical role; without it, asymptotic normality may fail. For instance, Example 3 in [47] demonstrates that the absence of condition (33) may result in the failure of asymptotic normality. Moreover, the matrix R_n can be chosen as $\Phi_n^{(11)}$ when the sequence $\{\varphi_k^{(1)}\}$ is predetermined. Alternatively, if the sequence $\varphi_n^{(1)}\varphi_n^{(1)\top}$ is stationary and ergodic with a positive definite covariance matrix, R_n can be set as $n\,E\big[\varphi_n^{(1)}\varphi_n^{(1)\top}\big]>0$, as suggested in [48].

Proof: Write $\hat{J}_n(\beta)$ in (13) as $\hat{J}_n(\beta)=\hat{J}_n(\beta_1,\beta_2)$ with $\beta_1\in\mathbb{R}^q$ and $\beta_2\in\mathbb{R}^{p-q}$, and let $\hat{\beta}_n=(\hat{\beta}_{1n}^\top,\hat{\beta}_{2n}^\top)^\top$ denote its minimizer. By Theorem 1, $\|\hat{\beta}_n-\theta\|\to 0$ a.s. Since each component of θ_{10} is nonzero, there exist $c>0$ and $N(\omega)<\infty$ such that $|\hat{\beta}_{1n}(j)|\ge c$ for all $j=1,\dots,q$ and all $n\ge N(\omega)$. Hence, the first-order optimality condition with respect to β_1 reads $\frac{\partial}{\partial\beta_1}\hat{J}_n(\hat{\beta}_{1n},\hat{\beta}_{2n})=0$, that is,

$$\begin{aligned}
&-2\sum_{k=1}^{n}\big(y_{k+1}-\hat{\beta}_{1n}^\top\varphi_k^{(1)}-\hat{\beta}_{2n}^\top\varphi_k^{(2)}\big)\varphi_k^{(1)}+2\lambda_{2n}P_{n,11}\hat{\beta}_{1n}\\
&+2\lambda_{2n}P_{n,12}\hat{\beta}_{2n}+\lambda_{1n}^{(2)}\,\mathrm{vec}\big(\mathrm{sign}(\hat{\beta}_{1n}(j))\,|\hat{\beta}_n^0(j)|^{-1}\big)\Big|_{j=1}^{q}=0. \qquad (35)
\end{aligned}$$

Using (1) and $\theta=[\theta_{10}^\top,0_{p-q}^\top]^\top$, we have $y_{k+1}=\theta_{10}^\top\varphi_k^{(1)}+w_{k+1}$. Substituting into (35) and rearranging yields

$$\begin{aligned}
&\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(1)\top}\Big(\big(I_q+\lambda_{2n}(\Phi_n^{(11)})^{-1}P_{n,11}\big)\hat{\beta}_{1n}-\theta_{10}\Big)\\
&=-\sum_{k=1}^{n}\hat{\beta}_{2n}^\top\varphi_k^{(2)}\varphi_k^{(1)}+\sum_{k=1}^{n}\varphi_k^{(1)}w_{k+1}-\lambda_{2n}P_{n,12}\hat{\beta}_{2n}\\
&\quad-\frac{1}{2}\lambda_{1n}^{(2)}\,\mathrm{vec}\big(\mathrm{sign}(\hat{\beta}_{1n}(j))\,|\hat{\beta}_n^0(j)|^{-1}\big)\Big|_{j=1}^{q}, \qquad (36)
\end{aligned}$$

Premultiplying (36) by $(\Phi_n^{(11)})^{-1}$ and then by $s_n^{-1}\upsilon_n^\top$ gives

$$\begin{aligned}
&s_n^{-1}\upsilon_n^\top\Big(\big(I_q+\lambda_{2n}(\Phi_n^{(11)})^{-1}P_{n,11}\big)\hat{\beta}_{1n}-\theta_{10}\Big)\\
&=-s_n^{-1}\upsilon_n^\top(\Phi_n^{(11)})^{-1}\Big(\sum_{k=1}^{n}\hat{\beta}_{2n}^\top\varphi_k^{(2)}\varphi_k^{(1)}+\lambda_{2n}P_{n,12}\hat{\beta}_{2n}\Big)\\
&\quad+s_n^{-1}\sum_{k=1}^{n}\upsilon_n^\top(\Phi_n^{(11)})^{-1}\varphi_k^{(1)}w_{k+1} \qquad (37)\\
&\quad-\frac{1}{2}\lambda_{1n}^{(2)}\,s_n^{-1}\upsilon_n^\top(\Phi_n^{(11)})^{-1}\,\mathrm{vec}\big(\mathrm{sign}(\hat{\beta}_{1n}(j))\,|\hat{\beta}_n^0(j)|^{-1}\big)\Big|_{j=1}^{q}.
\end{aligned}$$

For the first term on the right-hand side of (37), by Theorem 2 we have $\hat{\beta}_{2n}=0$ eventually, a.s., hence

$$\mathrm{P}\left(\lim_{n\to\infty}\sum_{k=1}^{n}\hat{\beta}_{2n}^\top\varphi_k^{(2)}\varphi_k^{(1)}+\lambda_{2n}P_{n,12}\hat{\beta}_{2n}=0\right)=1. \qquad (38)$$

For the last term on the right-hand side of (37), since $\hat{\beta}_{1n}\to\theta_{10}$ and $\hat{\beta}_n^0\to\theta_{10}$ while all entries of θ_{10} are nonzero, there exists a constant $c_5>0$ such that $|\hat{\beta}_n^0(j)|\ge c_5$ for $j=1,\dots,q$ when n is sufficiently large. Moreover, by the definition of s_n and the first condition in (33), we have $s_n^{-1}\lambda_{\min}(n)^{1/2}=O_p(1)$. Together with the assumption $\lambda_{1n}^{(2)}\,\lambda_{\min}(n)^{-1/2}\xrightarrow{P}0$, it follows that

$$\begin{aligned}
&\left|\lambda_{1n}^{(2)}s_n^{-1}\,\upsilon_n^\top\Big(\Phi_n^{(11)}\Big)^{-1}\,\mathrm{vec}\Big(\mathrm{sign}(\hat{\beta}_{1n}(j))\,|\hat{\beta}_n^0(j)|^{-1}\Big)\,\Big|_{j=1}^{q}\right|\\
&\le\frac{\lambda_{1n}^{(2)}}{\lambda_{\min}(n)^{1/2}}\cdot\frac{s_n^{-1}}{\lambda_{\min}(n)^{1/2}}\cdot\frac{q^{1/2}}{c_5}\ \xrightarrow{P}\ 0. \qquad (39)
\end{aligned}$$

Hence, combining (37), (38) and (39), we obtain

$$\begin{aligned}
&s_n^{-1}\,\upsilon_n^\top\Big(\big(I_q+\lambda_{2n}(\Phi_n^{(11)})^{-1}P_{n,11}\big)\hat{\beta}_{1n}-\theta_{10}\Big)\\
&=s_n^{-1}\sum_{k=1}^{n}\upsilon_n^\top\Big(\Phi_n^{(11)}\Big)^{-1}\varphi_k^{(1)}w_{k+1}\ +\ o_p(1). \qquad (40)
\end{aligned}$$

In view of (33) and (40), and by Slutsky's theorem [49], to

prove (34) it suffices to show that

$$s_n^{-1}\sum_{k=1}^{n} v_n^\top R_n^{-1}\varphi_k^{(1)} w_{k+1} \xrightarrow{\mathrm{d}} N(0,1). \tag{41}$$

Along the lines of [47], the desired limit (41) follows from the martingale central limit theorem of [50]. ∎

Remark 9: Theorem 3 immediately yields the asymptotic normality of the algorithm proposed in [18] as a special case. Let the estimator in [18] be denoted by $\bar{\beta}_n = (\bar{\beta}_{1n}^\top, \bar{\beta}_{2n}^\top)^\top$ with $\bar{\beta}_{1n} \in \mathbb{R}^q$. In particular, when $\lambda_{2n} = 0$, $\lambda_{1n}^{(1)} = 0$, and $\lambda_{1n}^{(2)}$ satisfies Assumption (A4) together with the condition $\lambda_{1n}^{(2)}/\lambda_{\min}(n)^{1/2} \xrightarrow{P} 0$, Theorem 3 implies that $\bar{\beta}_n$ enjoys the property $s_n^{-1} v_n^\top \left(\bar{\beta}_{1n} - \theta_{10}\right) \xrightarrow{\mathrm{d}} N(0,1)$.

C. Group effect

In many practical applications, predictors can be naturally partitioned into groups. As an illustrative example, consider an uplink massive MIMO system with M base station (BS) antennas and D users [51]: $y_m = \sum_{d=1}^{D} A_d x_{d,m} + n_m$, where y_m denotes the received frequency-domain signal at the m-th BS antenna, A_d is the known input matrix associated with the d-th user, and $x_{d,m}$ is the delay-domain channel vector of the d-th user to the m-th BS antenna. The goal of channel estimation is to recover $\{x_{d,m} : d = 1, \ldots, D,\, m = 1, \ldots, M\}$ from the set of received signals $\{y_m\}_{m=1}^{M}$.

Due to finite scattering in physical propagation, only a small fraction of the entries in $x_{d,m}$ are significant, while the majority are zero, implying that the channel vectors are sparse. Moreover, the channel vectors are naturally divided into groups according to the user index d. Within each group, the corresponding input variables tend to be highly correlated, for instance, the input matrix A_d may be low-rank. An effective channel estimation method should therefore satisfy two desirable properties: (i) eliminate insignificant paths, and (ii) automatically include all paths within a group once any single path in that group is selected (i.e., achieve group selection). Algorithm 1 fulfills both objectives: it performs variable selection while simultaneously selecting groups of correlated variables. Theorem 4 establishes that Algorithm 1 indeed possesses this group effect.

It is worth noting that a widely used method for enforcing group sparsity is the Group Lasso [52], [53], which applies an $L_{2,1}$ penalty to pre-defined groups of variables. While Group Lasso can effectively select or discard entire groups, it requires prior and exact knowledge of the grouping structure, and it cannot directly handle overlapping groups or adapt to correlation patterns that deviate from the pre-specified groups. Moreover, Group Lasso tends to shrink all coefficients within a group toward zero uniformly, which may lead to bias in estimating large coefficients compared with methods that balance L_1 sparsity and L_2 coupling. In contrast, Algorithm 1 does not rely on an explicit group partition: its L_2-coupling term automatically promotes similarity among highly correlated variables, allowing the group effect to emerge naturally from the data without strict prior grouping.

Theorem 4: Given the dataset $\{y_{k+1}, \varphi_k\}_{k=1}^{n}$, let $\hat{\beta}_n$ be the estimate produced by Algorithm 1 with the quadratic penalty term $\lambda_{2n}\beta^\top P_n \beta$, where $P_n > 0$. If $\hat{\beta}_n(i) > 0$ and $\hat{\beta}_n(j) > 0$ for some indices i and j, then

$$\left|(P_n\hat{\beta}_n)(i) - (P_n\hat{\beta}_n)(j)\right| \le \sqrt{\sum_{k=1}^{n} y_{k+1}^2}\sqrt{\sum_{k=1}^{n}\left(\varphi_k(i) - \varphi_k(j)\right)^2} \times \left(\frac{1}{\lambda_{2n}} + \frac{\lambda_{1n}^{(2)}}{2\lambda_{2n}^2 \min\left(|\hat{\beta}_n^0(i)|, |\hat{\beta}_n^0(j)|\right)^2}\right), \tag{42}$$

where $\hat{\beta}_n^0$ denotes the Step 1 estimate.

Furthermore, if P_n is diagonal on coordinates (i,j) with *equal* diagonal entries $p_{ii} = p_{jj} \ge \underline{p} > 0$, then (42) implies

$$|\hat{\beta}_n(i) - \hat{\beta}_n(j)| \le \frac{1}{\underline{p}}\sqrt{\sum_{k=1}^{n} y_{k+1}^2}\sqrt{\sum_{k=1}^{n}\left(\varphi_k(i) - \varphi_k(j)\right)^2} \times \left(\frac{1}{\lambda_{2n}} + \frac{\lambda_{1n}^{(2)}}{2\lambda_{2n}^2 \min\left(|\hat{\beta}_n^0(i)|, |\hat{\beta}_n^0(j)|\right)^2}\right). \tag{43}$$

Remark 10: Inequality (42) shows that the P_n-weighted coefficients $(P_n\hat{\beta}_n)(i)$ and $(P_n\hat{\beta}_n)(j)$ are close whenever $\sum_{k=1}^{n}\left(\varphi_k(i) - \varphi_k(j)\right)^2$ is small (e.g., after standardization, when the two regressors are highly correlated). If P_n is diagonal with $p_{ii} = p_{jj} \ge \underline{p} > 0$, or more generally block diagonal with the (i,j) block equal to αI_2 for some $\alpha \ge \underline{p}$, then (43) yields a direct bound on $|\hat{\beta}_n(i) - \hat{\beta}_n(j)|$. Both the L_2 term and the weighted L_1 term contribute to the group effect. However, in the pure L_1 case ($\lambda_{2n} = 0$), the bound in (43) degenerates as $\lambda_{2n} \downarrow 0$ and becomes uninformative; thus, this inequality alone does not establish a group effect when only the L_1 penalty is used.

Proof: Because $\hat{\beta}_n$ minimizes $\hat{J}_n(\beta)$ and $\hat{\beta}_n(i) > 0$, $\hat{\beta}_n(j) > 0$, the first order optimality conditions on the i-th and j-th coordinates read

$$-2\sum_{k=1}^{n}\left(y_{k+1} - \hat{\beta}_n^\top\varphi_k\right)\varphi_k(i) + \lambda_{1n}^{(2)}|\hat{\beta}_n^0(i)|^{-1} + 2\lambda_{2n}\left(P_n\hat{\beta}_n\right)(i) = 0, \tag{44}$$

$$-2\sum_{k=1}^{n}\left(y_{k+1} - \hat{\beta}_n^\top\varphi_k\right)\varphi_k(j) + \lambda_{1n}^{(2)}|\hat{\beta}_n^0(j)|^{-1} + 2\lambda_{2n}\left(P_n\hat{\beta}_n\right)(j) = 0. \tag{45}$$

Subtracting (45) from (44) gives

$$2\lambda_{2n}\Big((P_n\hat{\beta}_n)(i) - (P_n\hat{\beta}_n)(j)\Big) = -\lambda_{1n}^{(2)}\left(|\hat{\beta}_n^0(i)|^{-1} - |\hat{\beta}_n^0(j)|^{-1}\right) + 2\sum_{k=1}^{n}\left(y_{k+1} - \hat{\beta}_n^\top\varphi_k\right)\left(\varphi_k(i) - \varphi_k(j)\right).$$

Hence,

$$\lambda_{2n}\left|(P_n\hat{\beta}_n)(i) - (P_n\hat{\beta}_n)(j)\right| \le \frac{\lambda_{1n}^{(2)}}{2}\left|\,|\hat{\beta}_n^0(i)|^{-1} - |\hat{\beta}_n^0(j)|^{-1}\right| + \left|\sum_{k=1}^{n}\left(y_{k+1} - \hat{\beta}_n^\top\varphi_k\right)\left(\varphi_k(i) - \varphi_k(j)\right)\right|. \tag{46}$$

Since $\hat{\beta}_n$ is a minimizer, $\hat{J}_n(0_p) \ge \hat{J}_n(\hat{\beta}_n)$ implies $\sum_{k=1}^{n}\left(y_{k+1} - \hat{\beta}_n^\top\varphi_k\right)^2 + \lambda_{1n}^{(2)}\sum_{r=1}^{p}\frac{|\hat{\beta}_n(r)|}{|\hat{\beta}_n^0(r)|} + \lambda_{2n}\hat{\beta}_n^\top P_n\hat{\beta}_n \le \sum_{k=1}^{n} y_{k+1}^2$, which yields $\sum_{k=1}^{n}\left(y_{k+1} - \hat{\beta}_n^\top\varphi_k\right)^2 \le$

$\sum_{k=1}^{n} y_{k+1}^2$. Therefore, by Cauchy–Schwarz inequality,

$$\left|\sum_{k=1}^{n}\left(y_{k+1}-\hat{\beta}_n^\top\varphi_k\right)\left(\varphi_k(i)-\varphi_k(j)\right)\right| \le \sqrt{\sum_{k=1}^{n} y_{k+1}^2}\sqrt{\sum_{k=1}^{n}\left(\varphi_k(i)-\varphi_k(j)\right)^2}. \tag{47}$$

For the weighted L_1 term, by the mean value theorem there exists ξ between $|\hat{\beta}_n^0(i)|$ and $|\hat{\beta}_n^0(j)|$ such that

$$\left||\hat{\beta}_n^0(i)|^{-1}-|\hat{\beta}_n^0(j)|^{-1}\right| = \frac{\left||\hat{\beta}_n^0(i)|-|\hat{\beta}_n^0(j)|\right|}{\xi^2} \le \frac{|\hat{\beta}_n^0(i)-\hat{\beta}_n^0(j)|}{\left(\min\{|\hat{\beta}_n^0(i)|,|\hat{\beta}_n^0(j)|\}\right)^2}. \tag{48}$$

Plugging (47) and (48) into (46) gives

$$\begin{aligned}\lambda_{2n}&\left|(P_n\hat{\beta}_n)(i)-(P_n\hat{\beta}_n)(j)\right| \le \sqrt{\sum_{k=1}^{n} y_{k+1}^2}\sqrt{\sum_{k=1}^{n}\left(\varphi_k(i)-\varphi_k(j)\right)^2} \\ &+\frac{\lambda_{1n}^{(2)}}{2}\frac{|\hat{\beta}_n^0(i)-\hat{\beta}_n^0(j)|}{\left(\min\{|\hat{\beta}_n^0(i)|,|\hat{\beta}_n^0(j)|\}\right)^2}.\end{aligned} \tag{49}$$

Finally, applying the same argument to the Step 1 estimator $\hat{\beta}_n^0$ (the minimizer of J_n^0) yields $\lambda_{2n}\left|(P_n\hat{\beta}_n^0)(i)-(P_n\hat{\beta}_n^0)(j)\right| \le \sqrt{\sum_{k=1}^{n} y_{k+1}^2}\sqrt{\sum_{k=1}^{n}\left(\varphi_k(i)-\varphi_k(j)\right)^2}$. Combining this with (49) leads to (42). Moreover, if P_n is diagonal on the (i,j) coordinates with *equal* diagonal entries $p_{ii}=p_{jj}\ge \underline{p}>0$, then $\left|(P_n\hat{\beta}_n)(i)-(P_n\hat{\beta}_n)(j)\right| = \underline{p}\,|\hat{\beta}_n(i)-\hat{\beta}_n(j)|$, and (43) follows. ■

IV. Application to Identification of Linear Stochastic Systems under Self-Tuning Regulation Control

In this section, we apply Algorithm 1 to the sparse identification of a closed-loop system under a self-tuning regulator (STR) [54]. It is worth noting that in linear feedback control systems, the regressors are typically non-stationary and dependent [55], which poses additional challenges for parameter estimation.

We consider the following sparse ARX model:

$$\begin{aligned}y_{k+1} = a_1 y_k + \cdots + a_{n_y} y_{k+1-n_y} + b_1 u_k + \cdots \\ + b_{n_u} u_{k+1-n_u} + w_{k+1},\end{aligned} \tag{50}$$

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

For notational convenience, define

$$\begin{aligned}&A(z)=1-a_1z-\cdots-a_{n_y}z^{n_y},\\ &B(z)=b_1+b_2z+\cdots+b_{n_u}z^{n_u-1},\\ &\theta=\left[a_1,\ldots,a_{n_y},b_1,\ldots,b_{n_u}\right]^\top,\\ &\varphi_k=\left[y_k,\ldots,y_{k+1-n_y},u_k,\ldots,u_{k+1-n_u}\right]^\top.\end{aligned}$$

Let $\{y_k^*\}$ denote a deterministic and bounded reference signal. Within the framework of (50), our study addresses two main objectives: (i) applying STR control to ensure that the closed-loop system output y_k tracks the reference signal y_k^*; and (ii) achieving accurate identification of the system parameters—namely, correctly determining which coefficients are zero and consistently estimating the nonzero coefficients—under STR control.

Regarding the control phase, let $\theta_{L,n}=[a_{1,n},\ldots,a_{n_y,n},b_{1,n},\ldots,b_{n_u,n}]^\top$ denote the LS parameter estimate of the system. According to the Certainty Equivalence Principle [54], the adaptive control law can be expressed as

$$u_k^0=\frac{1}{b_{1,k}}\left(y_{k+1}^*+\left(b_{1,k}u_k-\theta_{L,k}^\top\varphi_k\right)\right). \tag{51}$$

In the identification phase, to ensure that the system's tracking performance is not degraded after introducing excitation, we adopt the *diminishing excitation* technique. Following [56], we add a vanishing perturbation term to the control law (51), leading to:

$$u_k=u_k^0+\frac{\nu_k}{r_{k-1}^{\bar{\varepsilon}/2}},\quad k\ge 1, \tag{52}$$

where $\{\nu_k\}$ is a bounded i.i.d. sequence with $E(\nu_k)=0$ and $E(\nu_k^2)=1$; $r_{k-1}=1+\sum_{i=1}^{k-1}\|\varphi_i\|^2$; $\bar{\varepsilon}\in\left(0,\frac{1}{2(t+1)}\right)$; and $t=\max\{n_y,n_u\}+n_y-1$.

We now introduce the assumptions on system (50) required for the subsequent stability and optimality analysis in Proposition 1:

(B1) The noise $\{w_k\}$ satisfies $\lim_{k\to\infty}\frac{1}{k}\sum_{j=1}^{k}w_j^2=R>0$ a.s.;

(B2) The system is minimum-phase, i.e., $B(z)\neq 0$ for all $|z|\le 1$;

(B3) $|a_{n_y}|+|b_{n_u}|\neq 0$.

Proposition 1: [56] Suppose that Assumptions (A1)–(A2) and (B1)–(B3) hold. Then, for the system (50) operating under the diminishing excitation control (51)–(52) based on the LS parameter estimate, we have $\limsup_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\left(\|u_i\|^2+\|y_i\|^2\right)<\infty$ a.s., and $\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\left(y_i-y_i^*\right)^2=R$ a.s., and the regressor φ_k satisfies the following excitation property:

$$\begin{aligned}&\lambda_{\max}(n)\triangleq\lambda_{\max}\Big(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\Big)=O(n),\\ &\lambda_{\min}(n)\triangleq\lambda_{\min}\Big(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\Big)\ge c\,n^{\,1-\bar{\varepsilon}(t+1)},\end{aligned}$$

for some $c>0$, which may depend on the sample path and the $\bar{\varepsilon}$ defined below (52).

Given the input–output data $\{y_{k+1},\varphi_k\}_{k=1}^n$ generated by (50)–(52), we apply Algorithm 1 to minimize the objective function (13) and obtain a sparse estimate of the system parameters: $\hat{\beta}_{L,n}=\left[\hat{\beta}_{L,n}(1),\ldots,\hat{\beta}_{L,n}(n_y+n_u)\right]^\top$. Define the true zero-parameter index set

$$\begin{aligned}D^*=\{&i: a_i=0 \text{ for } 1\le i\le n_y;\ b_{i-n_y}=0\\ &\text{for } n_y+1\le i\le n_y+n_u\},\end{aligned}$$

and the estimated zero-parameter index set

$$D_n^*=\left\{i:\hat{\beta}_{L,n}(i)=0,\ 1\le i\le n_y+n_u\right\}.$$

The following theorem establishes the convergence of D_n^*.

Theorem 5: Suppose that Assumptions (A1) and (B1)–(B3) hold, and consider Algorithm 1 with a positive

definite weight matrix P_n. Set $\lambda_{1n}^{(1)} = \lambda_{1n}^{(2)} = n^{1-\frac{5}{2}\bar{\varepsilon}(t+1)}$ and $\lambda_{2n}\lambda_{\max}(P_n) = n^{\tau}$, with $0 < \tau < 1 - \frac{5}{2}\bar{\varepsilon}(t+1)$ in Algorithm 1, and $\bar{\varepsilon} \in (0, \frac{2}{7(t+1)})$ in the controller (52). Then there exists a set Ω_0 with $P(\Omega_0) = 1$ such that for any $\omega \in \Omega_0$ there is an integer $N_0(\omega)$ satisfying $\widehat{D}_n^* = D^*$ for all $n \geq N_0(\omega)$.

Proof: By Theorem 2, it suffices to verify Assumptions (A3)(b) and (A4). First, since $\bar{\varepsilon} \in (0, \frac{1}{4})$, we have $\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} = O\Big(\frac{\sqrt{\log n}}{n^{\frac{1}{2}-\frac{3}{2}\bar{\varepsilon}(t+1)}}\Big) \xrightarrow[n\to\infty]{} 0$, and hence Assumption (A3)(b) is satisfied. Next, setting $\lambda_{2n}\lambda_{\max}(P_n) = n^{\tau}$ with $0 < \tau < 1 - \frac{5}{2}\bar{\varepsilon}(t+1)$, we obtain $\frac{\lambda_{1n}^{(1)}+\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)} = O\Big(\frac{1}{n^{\frac{3}{2}\bar{\varepsilon}(t+1)}}\Big) \to 0$, $\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}\frac{\lambda_{\max}(n)}{\lambda_{1n}^{(2)}} = O\Big(\frac{\log n}{n^{1-\frac{7}{2}\bar{\varepsilon}(t+1)}}\Big) \to 0$, $\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} = O\Big(\frac{\sqrt{\log n}}{n^{\frac{1}{2}-\frac{3}{2}\bar{\varepsilon}(t+1)}}\Big) \to 0$, and $\frac{\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)}{\lambda_{\min}(n)^2} = O\Big(\frac{1}{n^{\frac{1}{2}\bar{\varepsilon}(t+1)}}\Big) \to 0$. Therefore, $\{\lambda_{1n}^{(1)}, \lambda_{1n}^{(2)}, \lambda_{2n}P_n\}$ satisfy Assumption (A4), and the conclusion follows from Theorem 2. ∎

V. Simulation Study

This section presents three numerical simulations and one real-world example to evaluate the performance of Algorithm 1. The experiments cover two finite impulse response (FIR) systems, a linear feedback control system, and a vibrating flexible robot arm model from [57]. Since the objective function (13) is strictly convex, we employ the `CVX` toolbox for MATLAB to compute the minimizer in (14). All numerical tests are conducted in MATLAB R2025a on a Lenovo desktop (2.60 GHz, 32 GB RAM).

Example 1. This example illustrates the capability of Algorithm 1 in improving estimation accuracy and performing variable selection. Consider the FIR system $y_{k+1} = \theta^{\top}\varphi_k + w_{k+1}$, where $\theta \in \mathbb{R}^p$, the regressors $\{\varphi_k\}$ are i.i.d. p-dimensional Gaussian random vectors with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$, and the noise sequence $\{w_k\}$ is i.i.d. $N(0,1)$ and is independent of $\{\varphi_k\}$. To introduce collinearity among regressors, we consider two covariance structures:

(I) $\sigma_{ij} = \kappa^{|i-j|}$ with $\kappa = 0.7$;

(II) $\sigma_{ij} = \mathbb{I}(i=j) + \delta|i-j|^{-\nu}$ with $\delta = 0.5$ and $\nu = 1.5$.

We examine two settings: $(p,q) = (15,10)$ and $(p,q) = (40,15)$. The true parameter vector is $\theta = \big[1_{q/3}^{\top},\ 3\times 1_{q/3}^{\top},\ -2\times 1_{q/3}^{\top},\ 0_{p-q}^{\top}\big]^{\top}$. For comparison, we consider three methods: (i) the LS method [21]; (ii) the weighted L_1 method [18]; and (iii) Algorithm 1. For [18], we set $\lambda_n = n^{0.75}$. In Algorithm 1, we choose $\lambda_{1n}^{(1)} = \lambda_{1n}^{(2)} = n^{0.75}$ and $\lambda_{2n} = n^{0.15}$.

To assess estimation performance, we use the MSE. To evaluate variable selection, we compute the probability of correctly selecting zero coefficients (PCS), and the probability of incorrectly selecting nonzero coefficients (PIS). All experiments are repeated 50 times.

Tables I and II show the results under different settings. When the number of zero coefficients is small ($p = 15, q = 10$), the LS method yields a lower MSE but fails to produce sparse solutions. As the sample size and sparsity level increase, both Algorithm 1 and the method in [18] perform better, with Algorithm 1 consistently achieving lower MSE because its L_2 term helps balance variance and bias. Both methods produce sparse estimates; the method in [18] is slightly sparser but can erroneously zero nonzero coefficients.

TABLE I

Estimation error (MSE) and selection performance (PCS, PIS) under covariance structure (I): $\boldsymbol{\sigma_{ij} = \kappa^{|i-j|}}$, $\boldsymbol{\kappa = 0.7}$.

p	n	q	Method	MSE	PCS	PIS
15	100	10	LS	0.248 (0.428)	0	0
			Algorithm in [18]	0.393 (0.399)	0.997	0
			Algorithm 1	0.339 (0.380)	0.963	0
	200	10	LS	0.200 (0.250)	0	0
			Algorithm in [18]	0.294 (0.245)	0.997	0
			Algorithm 1	0.213 (0.152)	0.983	0
	400	10	LS	0.152 (0.167)	0	0
			Algorithm in [18]	0.176 (0.146)	1.000	0
			Algorithm 1	0.148 (0.120)	0.993	0
40	100	15	LS	1.301 (2.046)	0	0
			Algorithm in [18]	0.982 (1.180)	0.992	0.005
			Algorithm 1	0.888 (0.660)	0.970	0
	200	15	LS	0.680 (1.519)	0	0
			Algorithm in [18]	0.415 (0.423)	0.993	0
			Algorithm 1	0.383 (0.221)	0.985	0
	400	15	LS	0.339 (0.596)	0	0
			Algorithm in. [18]	0.271 (0.357)	0.999	0.001
			Algorithm 1	0.256 (0.169)	0.992	0

TABLE II

Estimation accuracy (MSE) and selection performance (PCS, PIS) under covariance structure (II): $\boldsymbol{\sigma_{ij} = \mathbb{I}(i=j) + \delta|i-j|^{-\nu}}$ with $\boldsymbol{\delta = 0.5}$ and $\boldsymbol{\nu = 1.5}$.

p	n	q	Method	MSE	PCS	PIS
15	100	10	LS	0.343 (0.413)	0.50	0
			Algorithm in [18]	0.409 (0.498)	0.993	0.002
			Algorithm 1	0.346 (0.323)	0.977	0
	200	10	LS	0.170 (0.232)	0.52	0
			Algorithm in [18]	0.211 (0.287)	0.997	0
			Algorithm 1	0.179 (0.197)	0.990	0
	400	10	LS	0.069 (0.112)	0.52	0
			Algorithm in [18]	0.090 (0.091)	1.000	0
			Algorithm 1	0.061 (0.081)	1.000	0
40	100	15	LS	1.754 (2.729)	0.494	0
			Algorithm in [18]	0.844 (0.946)	0.989	0.007
			Algorithm 1	0.592 (0.418)	0.990	0
	200	15	LS	0.240 (0.350)	0.496	0
			Algorithm in [18]	0.191 (0.213)	0.997	0
			Algorithm 1	0.177 (0.139)	0.998	0
	400	15	LS	0.224 (0.292)	0.498	0
			Algorithm in [18]	0.162 (0.181)	0.994	0
			Algorithm 1	0.140 (0.123)	1.000	0

Example 2. This example illustrates the group selection property of Algorithm 1. Consider a multi-group FIR model $y_{k+1} = \sum_{g=1}^{G}\theta_g^{\top}\varphi_k^{(g)} + w_{k+1}$, with $G = 10$ groups and $\sum_{g=1}^{G} p_g = 100$. The true coefficient vector is randomly generated so that all coefficients in each active group have the same sign, while coefficients in inactive groups are set to zero. For each group g, the regressor is generated as $\varphi_k^{(g)} = \sqrt{\rho}\, b_k^{(g)}\mathbf{1}_{p_g} + \sqrt{1-\rho}\,\epsilon_k^{(g)}$, where $\rho = 0.99$, $b_k^{(g)} \sim \mathcal{N}(0,1)$,

and $\epsilon_k^{(g)}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Regressors from different groups are mutually independent. The noise sequence $\{w_k\}$ is i.i.d. $\mathcal{N}(0,1.5^2)$ and is independent of all regressors. We set $N = 150$ and run 100 Monte Carlo trials, comparing Algorithm 1 with: (i) group LASSO with known and correct group partition; (ii) group LASSO without group information, where variables are randomly regrouped; (iii) weighted L_1 [18]; and (iv) LASSO [7]. All methods use the regularization level $N\sqrt{\log N/N}$ as in (10) for a fair comparison.

Fig. 1 presents two performance measures: the top panel shows boxplots of selection accuracy, and the bottom panel displays the empirical cumulative distribution functions (ECDFs) of the L_2 estimation errors. With highly correlated variables within each active group, Algorithm 1 tends to retain all variables in active groups, yielding higher and more stable selection accuracy with smaller estimation errors. Other methods may select only a subset of variables within some groups, reducing the stability of support recovery.

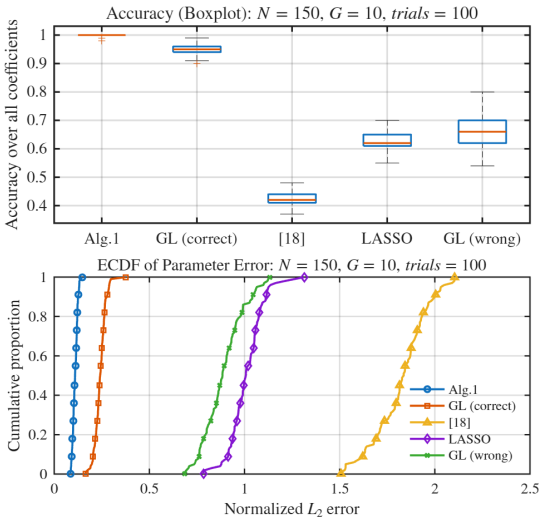

Fig. 1. Comparative results in Example 2: boxplots of selection accuracy (top) and ECDFs of estimation errors (bottom).

Example 3. This example demonstrates Algorithm 1 on a sparse linear feedback control system with non-stationary and dependent regressors, and examines the impact of P_n selection. Consider the ARX model (1) with

$$\varphi_k = \left[y_k,\, y_{k-1},\ldots,y_{k+1-n_y},\, u_k,\, u_{k-1},\ldots,u_{k+1-n_u}\right]^\top,$$

where the noise $\{w_k\}$ is i.i.d. $N(0,0.5)$. We set $N = 1000$ samples and repeat each experiment for 100 trials. The true parameter $\theta = [a^\top, b^\top]^\top$ is sparse: the AR part $a \in \mathbb{R}^{45}$ has five nonzeros (at indices $\{1,2,23,24,45\}$ with magnitudes 0.52, -0.17, 0.10, 0.04, and 0.18), and the input part $b \in \mathbb{R}^{12}$ has four nonzeros (at indices $\{1,4,6,9\}$ with magnitudes 0.80, 0.50, 0.35, and 0.25). This setting satisfies Assumptions (B1)–(B3). The reference signal is $y^*_{k+1} = 25\sin(\frac{k}{100}) - 2\cos(\frac{k}{200})$, $k \geq 0$. We use the certainty-equivalence diminishing excitation as in (52). Simulation parameters are set as follows: $\bar{\varepsilon} = 0.02$ and $\nu_k \sim \mathrm{U}(-\sqrt{3},\sqrt{3})$. A representative trajectory and the corresponding control input are shown in Fig. 2, which demonstrates good control performance.

For each trial and along a growing sample grid $n \in \{100,\ldots,1000\}$, we compare seven estimators: five variants of Algorithm 1 (Alg. 1) with different quadratic penalties, weighted L_1 [18], and LS. The five variants of Alg. 1 correspond to different choices of P_n: (I) $P_n = I_{57}$; (II) $P_n = \mathrm{diag}\{0.5I_{45},\, 1.8I_{12}\}$; (III) $P_n = D^\top D$, where $D \in \mathbb{R}^{(p-1)\times p}$ is the first-order difference matrix with $D_{i,i} = -1$ and $D_{i,i+1} = 1$ [41]; (IV) $P_n = \mathrm{blkdiag}(L_y, L_u)$ with $L_y = n_y I_{n_y} - \mathbf{1}_{n_y}\mathbf{1}_{n_y}^\top$ and $L_u = n_u I_{n_u} - \mathbf{1}_{n_u}\mathbf{1}_{n_u}^\top$. We then set $P_n \leftarrow (1-\alpha)\,P_n/\lambda_{\max}(P_n) + \alpha I$ with $\alpha = 0.7$, following standard Laplacian-based regularization [42]; (V) $P_n = (1-\rho)K_n + \rho\mu I_p$, where $K_n = \frac{1}{n}\sum_{k=1}^n \varphi_k\varphi_k^\top$, $\mu = \frac{1}{p}\mathrm{tr}(K_n)$, and $\rho = 0.4$. The resulting matrix is symmetrized and scaled to unit spectral radius, following the shrinkage approach of [43]. These constructions ensure that P_n is positive definite.

We evaluate performance by the mean-squared error $\mathrm{MSE}(\hat{\theta}) = \|\hat{\theta}-\theta\|_2^2/p$. Fig. 3 shows the median and interquartile range (IQR) of the MSE as the sample size n increases. Alg. 1 yields lower median errors and smaller variation than both weighted L_1 and LS. The boxplot at $n = N$ in Fig. 4 further confirms the advantage of Alg. 1.

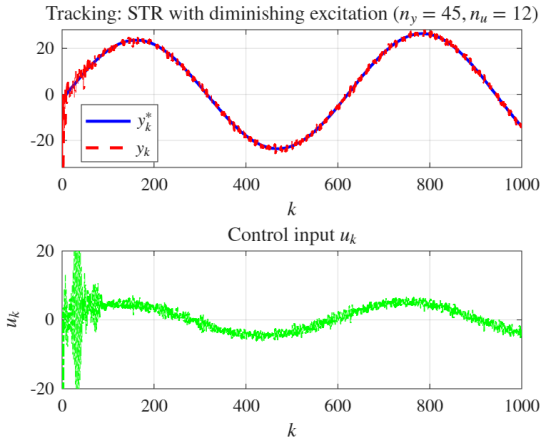

Fig. 2. Tracking results in Example 3: reference and actual outputs (top), and control input (bottom).

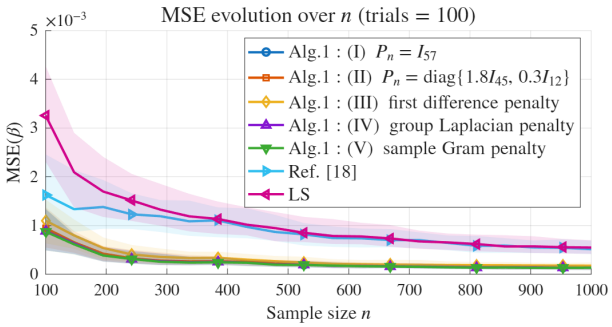

Fig. 3. Median and interquartile range bands of MSE over sample size n across different algorithms.

Example 4. A high-fidelity robot model is essential for precise positioning and for minimizing tracking errors in robotic applications. Following the setup in [57], we use the vibrating flexible robot arm dataset, with experimental analyses also reported in [30], [58]. The dataset contains 40,960 samples at 500 Hz. The input is the driving torque and the output

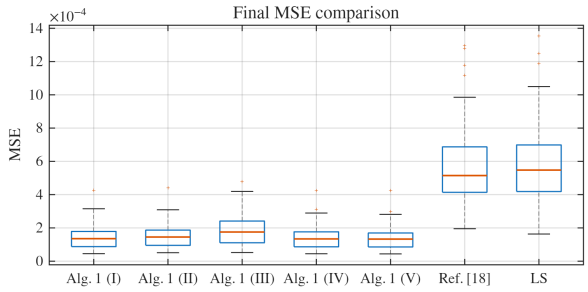

Fig. 4. Final MSE comparison at $\boldsymbol{n} = \boldsymbol{N}$ for different algorithms. Variants of Alg. 1 are denoted by (I)–(V), corresponding to different $\boldsymbol{P_n}$.

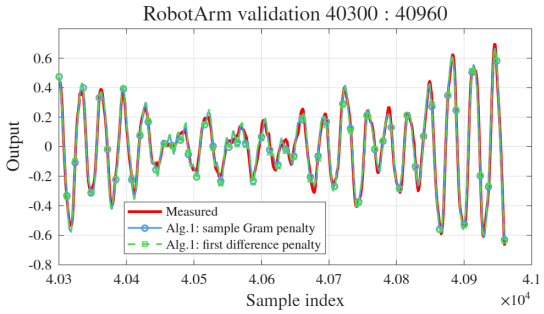

Fig. 5. Measured output v.s. the estimated output of two best models.

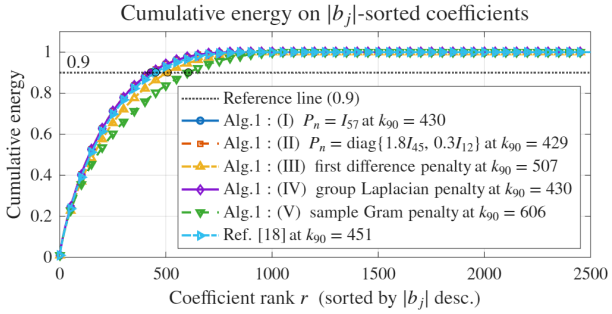

Fig. 6. Cumulative energy curves of absolute-sorted coefficients for six algorithms.

is the acceleration measured at the arm's tip. We compare Algorithm 1 with the methods in [18], [30], [58]. Since the true system is unknown, model quality is assessed via a hold-out validation: the first 1:7000 data points are used for estimation, and the interval 10,000:40,960 is reserved for validation. All identified models are of order 2500. To evaluate performance, we use the FIT index: $\mathrm{FIT} = 100\left(1 - \frac{\|y-\hat{y}\|}{\|y-\mathrm{mean}(y)\|}\right)$, where y is the measured output and $\hat{y}$ is the simulated output. A FIT value closer to 100% indicates higher simulation accuracy.

Table III reports the validation FIT for five Alg.1 variants (I)–(V), with configurations identical to Example 3. Among them, the sample-Gram variant (V) attains the highest FIT (85.209%). For reference, the 80.100% and 79.900% FITs are quoted from the original papers [30], [58]; all Alg.1 variants outperform these baselines. Moreover, Fig. 5 compares the measured output with the two best Alg.1 variants (V) and (III) on the window $[40{,}300,\ 40{,}960]$, confirming the high estimation accuracy. Fig. 6 shows that all six methods produce sparse models. Among them, the curve of Alg.1 (V) lies farthest to the right (i.e., it needs more coefficients to reach 90% energy). This means Alg.1 (V) keeps some medium–size coefficients instead of discarding them. These coefficients capture small effects and improve data fitting. Fig. 7 displays the top–100 coefficients. Alg.1 (V) concentrates most of them in the lag range 0–600, while the other methods either miss some in this range or spread weight to longer lags. This focus within 0–600 matches the system's short memory.

TABLE III
FIT OF DIFFERENT ALGORITHMS ON EXAMPLE 4.

Algorithm	FIT
Alg.1: (I) $P_n = I_{57}$	85.140%
Alg.1: (II) $P_n = \mathrm{diag}\{1.8\,I_{45},\ 0.3\,I_{12}\}$	85.138%
Alg.1: (III) first-difference penalty	85.169%
Alg.1: (IV) group Laplacian penalty	85.140%
Alg.1: (V) sample Gram penalty	85.209%
[18]	83.988%
[58]	80.100%
[30]	79.900%

VI. CONCLUSION

This paper presents a refined identification framework that simultaneously addresses collinearity and promotes sparsity, with theoretical guarantees of almost sure parameter convergence, almost sure set convergence, and asymptotic normality under non-i.i.d. and non-stationary observations. The proposed approach integrates weighted L_1 and L_2 regularization: the weighted L_1 term induces sparsity, while the L_2 term enhances stability, improves estimation accuracy, and enables group selection. Its effectiveness has been demonstrated in sparse parameter identification for linear feedback control systems.

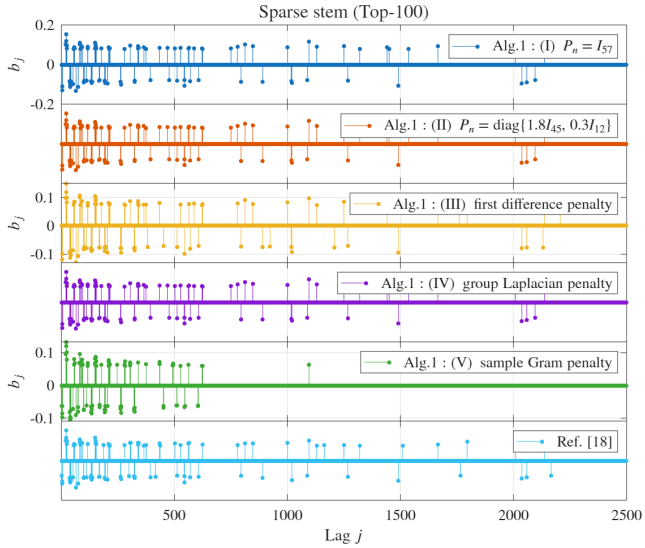

Fig. 7. Sparse stem plots of the top-100 coefficients for six algorithms.

Future research directions include extending the approach to stochastic sparse systems in high-dimensional settings in which $p = p(n)$, and developing recursive algorithms suitable for real-time identification and control design.

REFERENCES

[1] I. Kropp, A. P. Nejadhashemi, and K. Deb, "Benefits of sparse population sampling in multi-objective evolutionary computing for large-scale sparse optimization problems," *Swarm Evol. Comput.*, vol. 69, Jun. 2022, Art. no. 108533.

[2] E. W. Bai, C. M. Cheng, and W. X. Zhao, "Variable selection of high-dimensional non-parametric nonlinear systems by derivative averaging to avoid the curse of dimensionality," *Automatica*, vol. 101, pp. 138–149, 2019.

[3] G. Fatima, P. Babu, and P. Stoica, "Two new algorithms for maximum likelihood estimation of sparse covariance matrices with applications to graphical modeling," *IEEE Trans. Signal Process.*, vol. 72, no. 8, pp. 2143–2156, Apr. 2024.

[4] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Ann. Stat.*, vol. 37, no. 4, pp. 1733–1751, 2009.

[5] Y. J. Tian and Y. Q. Zhang, "A comprehensive survey on regularization strategies in machine learning," *Inf. Fusion*, vol. 80, pp. 146–166, 2022.

[6] D. Bertsimas, J. Pauphilet, and B. Van Parys, "Sparse regression," *Statist. Sci.*, vol. 35, no. 4, pp. 555–578, 2020.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[8] J. Q. Fan and R. Z. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[9] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[10] H. Zou, "The adaptive LASSO and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[11] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Stat.*, vol. 38, no. 2, pp. 894–942, 2010.

[12] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Stat.*, vol. 43, no. 4, pp. 1535–1567, 2015.

[13] S. Song and P. J. Bickel, "Large vector autoregressions," *J. Amer. Statist. Assoc.*, vol. 106, no. 496, pp. 1376–1387, 2011.

[14] W. Wu, A. Shojaie, and G. Michailidis, "Performance bounds for parameter estimates of sparse vector autoregressive models," *Biometrika*, vol. 103, no. 4, pp. 899–916, 2016.

[15] R. P. Masini, M. C. Medeiros, and E. F. Mendes, "Machine learning advances for time series forecasting," *J. Econ. Surv.*, vol. 37, no. 1, pp. 76–111, 2023.

[16] B. Babadi, N. Kalouptsidis, and V. Tarokh, "SPARLS: The sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, 2010.

[17] B. Dumitrescu, A. Onose, P. Helin, and I. Tabus, "Greedy sparse RLS," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2194–2207, 2012.

[18] W. X. Zhao, G. Yin, and E.-W. Bai, "Sparse system identification for stochastic systems with general observation sequences," *Automatica*, vol. 121, 2020, Art. no. 109162.

[19] J. Parsa, C. R. Rojas, and H. Hjalmarsson, "Transformation of regressors for low coherent sparse system identification," *IEEE Trans. Autom. Control*, vol. 69, no. 5, pp. 2947–2962, May 2024.

[20] ——, "Balancing application relevant and sparsity revealing excitation in input design," *IEEE Trans. Autom. Control*, vol. 70, no. 3, pp. 1890–1897, Mar. 2025.

[21] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*. Berlin, Germany: Springer, 2012.

[22] S. K. Guharay, G. S. Thakur, F. J. Goodman, S. L. Rosen, and D. Houser, "Analysis of non-stationary dynamics in the financial system," *Econ. Lett.*, vol. 121, no. 3, pp. 454–457, 2013.

[23] F. Sattler, S. Wiedemann, K. R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, 2019.

[24] S. B. Wu, C. X. Wang, M. M. Alwakeel, and X. H. You, "A general 3-d non-stationary 5g wireless channel model," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3065–3078, Jul. 2017.

[25] L. B. Cao, "Beyond iid: Non-iid thinking, informatics, and learning," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 5–17, Jul. 2022.

[26] Y. X. Fu and W. X. Zhao, "Support recovery and parameter identification of multivariate arma systems with exogenous inputs," *SIAM J. Control Optim.*, vol. 61, no. 3, pp. 1835–1860, 2023.

[27] J. Guo, Y. Wang, Y. Zhao, and J.-F. Zhang, "Sparse parameter identification for stochastic systems based on l_γ regularization," *SIAM J. Control Optim.*, vol. 62, no. 6, pp. 2884–2909, 2024.

[28] J. Q. Fan and J. C. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Statist. Soc. Ser. B Stat. Methodol.*, vol. 70, no. 5, pp. 849–911, 2008.

[29] G. Pillonetto, T. S. Chen, A. Chiuso, G. De Nicolao, and L. Ljung, *Regularized System Identification: Learning Dynamic Models from Data*. Cham, Switzerland: Springer, 2022.

[30] G. Pillonetto, F. Dinuzzo, T. S. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[31] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston & Sons, 1977.

[32] D. L. Phillips, "A technique for the numerical solution of certain integral equations of the first kind," *J. ACM*, vol. 9, no. 1, pp. 84–97, 1962.

[33] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[34] B. Q. Mu and T. S. Chen, "On asymptotic optimality of cross-validation based hyper-parameter estimators for kernel-based regularized system identification," *IEEE Trans. Autom. Control*, vol. 69, no. 1, pp. 593–608, Jan. 2024.

[35] Y. Ju, B. Q. Mu, L. Ljung, and T. S. Chen, "Asymptotic theory for regularized system identification part i: Empirical bayes hyper-parameter estimator," *IEEE Trans. Autom. Control*, vol. 68, no. 12, pp. 7297–7312, Dec. 2023.

[36] T. L. Lai and C. Z. Wei, "least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *Ann. Stat.*, vol. 10, no. 1, pp. 154–166, 1982.

[37] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: identifying influential data and sources of collinearity*. New York: Wiley, 1980.

[38] W. Cao and G. Pillonetto, "Dealing with collinearity in large-scale linear system identification using gaussian regression," *Automatica*, vol. 160, 2024, Art. no. 111708.

[39] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Márquez, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 27–46, 2013.

[40] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.

[41] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

[42] D. Bianchi, D. Evangelista, S. Aleotti, M. Donatelli, E. L. Piccolomini, and W. Li, "A data-dependent regularization method based on the graph laplacian," *SIAM J. Sci. Comput.*, vol. 47, no. 2, pp. C369–C398, 2025.

[43] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[44] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[46] T. L. Lai and C. Z. Wei, "On the concept of excitation in least squares identification and adaptive control," *Stochastics*, vol. 16, no. 3-4, pp. 227–254, 1986.

[47] T. L. Lai and H. Robbins, "Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes," *Z. Wahrsch. Verw. Gebiete*, vol. 56, pp. 329–360, 1981.

[48] L. T. Zhang and L. Guo, "Adaptive identification with guaranteed performance under saturated observation and nonpersistent excitation," *IEEE Trans. Autom. Control*, vol. 69, no. 3, pp. 1584–1599, Mar. 2024.

[49] J. Shao, *Mathematical Statistics*. New York, NY: Springer, 2003.

[50] A. Dvoretzky, "Asymptotic normality for sums of dependent random variables," in *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, vol. 2. Berkeley, CA: Univ. California Press, 1972, pp. 513–535.

[51] X. T. Cheng, K. Xu, J. J. Sun, and S. Q. Li, "Adaptive grouping sparse bayesian learning for channel estimation in non-stationary uplink

massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4184–4198, 2019.
[52] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Statist.*, vol. 22, no. 2, pp. 231–245, 2013.
[53] T. T. Cai, A. R. Zhang, and Y. Zhou, "Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference," *IEEE Trans. Inf. Theory*, vol. 68, no. 9, pp. 5975–6002, 2022.
[54] K. J. Åström and B. Wittenmark, "On self tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, 1973.
[55] D. W. Huang and L. Guo, "Estimation of nonstationary ARMAX models based on the hannan-rissanen method," *Ann. Stat.*, vol. 18, no. 4, pp. 1729–1756, 1990.
[56] L. Guo and H. F. Chen, "The astrom-wittenmark self-tuning regulator revisited and els-based adaptive trackers," *IEEE Trans. Autom. Control*, vol. 36, no. 7, pp. 802–812, Jul. 1991.
[57] D. E. Torfs, R. Vuerinckx, J. Swevers, and J. Schoukens, "Comparison of two feedforward design methods aiming at accurate trajectory tracking of the end point of a flexible robot arm," *IEEE Trans. Control Syst. Technol.*, vol. 6, no. 1, pp. 2–14, Jan. 1998.
[58] G. C. Calafiore, C. Novara, and M. Taragna, "Leading impulse response identification via the Elastic Net criterion," *Automatica*, vol. 80, pp. 75–87, 2017.

APPENDIX

We first present a useful technical result.

Lemma 4: Let $A, B \in \mathbb{R}^{p\times p}$ be positive definite matrices. If $A^2 \le B^2$, then $\|B^{-1}A\| \le 1$.

Proof: For any nonzero vector $x \in \mathbb{R}^p$, let $y = B^{-1}x$. Then $\|AB^{-1}x\|^2 = \|Ay\|^2 = y^\top A^2 y \le y^\top B^2 y = x^\top x = \|x\|^2$. Since $B^{-1}A$ is the adjoint of AB^{-1}, it follows that $\|B^{-1}A\| = \|AB^{-1}\| = \sup_{x\neq 0}\frac{\|AB^{-1}x\|}{\|x\|} \le 1$. ∎

Proof of Lemma 3. By the definition of $\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)$, it follows that

$$J_n\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n),\alpha_{1n},\alpha_{2n},\eta_n\big) \le J_n(\theta,\alpha_{1n},\alpha_{2n},\eta_n). \tag{53}$$

Using (1) and (15), and noting that $\theta(j) = 0$ for $j = q+1,\dots,p$, we obtain

$$\begin{aligned} &J_n(\theta,\alpha_{1n},\alpha_{2n},\eta_n)\\ &= \sum_{k=1}^{n} w_{k+1}^2 + \alpha_{1n}\sum_{j=1}^{q}\eta_n(j)\,|\theta(j)| + \alpha_{2n}\,\theta^\top P_n\theta. \end{aligned} \tag{54}$$

and

$$\begin{aligned} &J_n\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n),\alpha_{1n},\alpha_{2n},\eta_n\big)\\ &=\sum_{k=1}^{n}\big(y_{k+1}-\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)^\top\varphi_k\big)^2\\ &\quad+\alpha_{1n}\sum_{j=1}^{p}\eta_n(j)\,\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)\big| + \alpha_{2n}\,\beta_n^\top P_n\,\beta_n\\ &=\sum_{k=1}^{n} w_{k+1}^2+\sum_{k=1}^{n}\big((\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta)^\top\varphi_k\big)^2\\ &\quad-2\,(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta)^\top\sum_{k=1}^{n}\varphi_k w_{k+1}\\ &\quad+\alpha_{1n}\sum_{j=1}^{p}\eta_n(j)\,\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)\big|\\ &\quad+\alpha_{2n}\,\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)^\top P_n\,\beta_n(\alpha_{1n},\alpha_{2n},\eta_n), \end{aligned} \tag{55}$$

Noting that $\alpha_{1n}\sum_{j=q+1}^{p}\eta_n(j)\,\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)\big| \ge 0$, and using (55)–(54), we have

$$\begin{aligned} &J_n\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n),\alpha_{1n},\alpha_{2n},\eta_n\big) - J_n(\theta,\alpha_{1n},\alpha_{2n},\eta_n)\\ &\ge\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big)^\top\sum_{k=1}^{n}\varphi_k\varphi_k^\top\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big)\\ &\quad+\alpha_{2n}\left[\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)^\top P_n\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta^\top P_n\theta\right]\\ &\quad-2\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big)^\top\sum_{k=1}^{n}\varphi_k w_{k+1}\\ &\quad+\alpha_{1n}\sum_{j=1}^{q}\eta_n(j)\left(\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)\big|-|\theta(j)|\right)\\ &\triangleq M_n^{(1)}+M_n^{(2)}-2M_n^{(3)}+M_n^{(4)}. \end{aligned} \tag{56}$$

Now we estimate $M_n^{(i)}$, $i = 1,\dots,4$, separately. Let

$$\delta_n = \left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top+\alpha_{2n}P_n\right)^{\frac{1}{2}}\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big). \tag{57}$$

First, by direct calculation, it follows $M_n^{(1)}+M_n^{(2)} = \delta_n^\top\delta_n - 2\alpha_{2n}\,\theta^\top P_n\big(\theta-\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)\big)$. Noting that $\big|\theta^\top P_n(\theta-\beta_n(\alpha_{1n},\alpha_{2n},\eta_n))\big| \le \lambda_{\max}(P_n)\|\theta\|\times\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|$, we then have

$$\begin{aligned} &M_n^{(1)}+M_n^{(2)}\\ &\ge\delta_n^\top\delta_n - 2\alpha_{2n}\lambda_{\max}(P_n)\,\|\theta\|\,\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|. \end{aligned} \tag{58}$$

Second, noting that $\sum_{k=1}^{n}\varphi_k\varphi_k^\top \le \sum_{k=1}^{n}\varphi_k\varphi_k^\top+\alpha_{2n}P_n$, by Lemma 4 we have $\left\|\big(\sum_{k=1}^{n}\varphi_k\varphi_k^\top+\alpha_{2n}P_n\big)^{-\frac{1}{2}}\big(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\big)^{\frac{1}{2}}\right\| \le 1$. Hence,

$$\begin{aligned} |M_n^{(3)}| &= \left|(\beta_n-\theta)^\top\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top+\alpha_{2n}P_n\right)^{\frac{1}{2}}\right.\\ &\qquad\left.\times\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top+\alpha_{2n}P_n\right)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k w_{k+1}\right|\\ &\le\left\|(\beta_n-\theta)^\top\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top+\alpha_{2n}P_n\right)^{\frac{1}{2}}\right\|\\ &\qquad\times\left\|\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\right)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k w_{k+1}\right\|\\ &= O\Big(\sqrt{\log\lambda_{\max}(n)}\Big)\,\|\delta_n\|, \end{aligned}$$

so there exists a constant $c_1 > 0$ such that, for all sufficiently large n,

$$M_n^{(3)} \ge -\frac{c_1}{2}\sqrt{\log\lambda_{\max}(n)}\,\|\delta_n\|. \tag{59}$$

Last, for $M_n^{(4)}$, by the Cauchy–Schwarz inequality,

$$\begin{aligned} |M_n^{(4)}| &\le \alpha_{1n}\sum_{j=1}^{q}\eta_n(j)\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)-\theta(j)\big|\\ &\le \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}\,\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|. \end{aligned} \tag{60}$$

Substituting (53), (58), (59), and (60) into (56) yields

$$0 \geq \|\delta_n\|^2 - \alpha_{2n}\lambda_{\max}(P_n)\|\theta\|\,\|\beta_n-\theta\| - c_1\sqrt{\log\lambda_{\max}(n)}\,\|\delta_n\| - \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}\,\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|. \quad (61)$$

By solving the quadratic inequality (61) with respect to $\|\delta_n\|$ and using $(a+b)^2 \leq 2a^2+2b^2$, we obtain

$$\|\delta_n\|^2 \leq c_1^2\log\lambda_{\max}(n) + 2\alpha_{2n}\lambda_{\max}(P_n)\|\theta\|\,\|\beta_n-\theta\| + 2\alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}\,\|\beta_n-\theta\|. \quad (62)$$

where, for brevity, β_n denotes $\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)$. Moreover, recalling (57) and using

$$\lambda_{\min}\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top + \alpha_{2n}P_n\right) \geq \lambda_{\min}(n) + \alpha_{2n}\lambda_{\min}(P_n),$$

we have $\|\delta_n\|^2 \geq \left(\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)\right)\|\beta_n-\theta\|^2$. Combining these yields

$$\left(\|\beta_n-\theta\| - \frac{\alpha_{2n}\lambda_{\max}(P_n)\|\theta\| + \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\right)^2 \leq \left(\frac{\alpha_{2n}\lambda_{\max}(P_n)\|\theta\| + \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\right)^2 + \frac{c_1^2\log\lambda_{\max}(n)}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}$$

$$\leq \left(\frac{\alpha_{2n}\lambda_{\max}(P_n)\|\theta\| + \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)} + \sqrt{\frac{c_1^2\log\lambda_{\max}(n)}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}}\right)^2.$$

Consequently,

$$\|\beta_n-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}} + \frac{\alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}+\alpha_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\right),$$

which completes the proof. ■

Jian Guo received the B.S. degree in Mathematics from Xi'an Jiaotong University, Xi'an, China, in 2019, and the Ph.D. degree in system analysis and integration from the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS), Beijing, China, in 2024. He is currently a postdoc researcher at the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. His research interests include sparse identification, control of multi-agent systems, and dynamic stochastic variational inequalities.

Ying Wang (IEEE Member) received the B.S. degree in mathematics from Wuhan University, Wuhan, China, in 2017, and the Ph.D. degree in system theory from the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS), Beijing, China, in 2022. She is currently a postdoc researcher at Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden and the AMSS, CAS. Her research interests include parameter identification and adaptive control of quantized systems. She was a recipient of the Special Prize of the Presidential Scholarship of Chinese Academy of Sciences in 2022, and the Outstanding Doctoral Dissertation Award of the Chinese Association of Automation in 2024.

Yanlong Zhao (Senior Member, IEEE) received the B.S. degree in mathematics from Shandong University, Jinan, China, in 2002, and the Ph.D. degree in systems theory from the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS), Beijing, China, in 2007. Since 2007, he has been with the AMSS, CAS, where he is currently a full Professor and a Vice Director of State Key Laboratory of Mathematical Sciences. His research interests include identification and control of quantized systems, networked systems, information theory and modeling of financial systems.

He received the second prize of the State Natural Science Award of China in 2015. He has been a Deputy Editor-in-Chief Journal of Systems and Science and Complexity, an Associate Editor of Automatica, SIAM Journal on Control and Optimization, and IEEE Transactions on Systems, Man and Cybernetics: Systems. He served as a Vice President of Asian Control Association and IEEE CSS Beijing Chapter, and is now a Vice President of Chinese Association of Automation (CAA), Chair of Technical Committee on Control Theory (TCCT), CAA and member of IFAC TC 1.1 Modeling, Identification & Signal Processing.

Ji-Feng Zhang (IEEE Fellow) received the B.S. degree in mathematics from Shandong University, China, in 1985, and the Ph.D. degree from the Institute of Systems Science, Chinese Academy of Sciences (CAS), China, in 1991. Now he is with the School of Automation and Electrical Engineering, Zhongyuan University of Technology; and the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, CAS. His current research interests include system modeling, adaptive control, stochastic systems, and multi-agent systems.

He is an IEEE Fellow, IFAC Fellow, CAA Fellow, CSIAM Fellow, member of the European Academy of Sciences and Arts, and Academician of the International Academy for Systems and Cybernetic Sciences. He received the Second Prize of the State Natural Science Award of China in 2010 and 2015, respectively. He was a Vice-President of the Chinese Association of Automation, the Chinese Mathematical Society and the Systems Engineering Society of China. He was a Vice-Chair of the IFAC Technical Board, member of the Board of Governors, IEEE Control Systems Society; Convenor of Systems Science Discipline, Academic Degree Committee of the State Council of China. He served as Editor-in-Chief, Deputy Editor-in-Chief or Associate Editor for more than 10 journals, including Science China Information Sciences, IEEE Transactions on Automatic Control and SIAM Journal on Control and Optimization etc