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*— Sparse identification of stochastic systems faces regressor collinearity and dependent, possibly non-stationary observations generated by system dynamics and feedback. This paper proposes a two-stage adaptive identification method that combines a weighted L_1 penalty for sparse recovery with quadratic P_n-regularization for stable estimation under collinearity. Our theory covers two ill-conditioning regimes with distinct identification targets. First, under increasingly severe collinearity, we address recovery of the original parameter when the condition number of the information matrix may diverge while sufficient information for identification is still retained. Under excitation conditions allowing non-i.i.d. and nonstationary observations, we prove finite-step identification of all zero coefficients with probability one, almost sure convergence, and asymptotic normality of the nonzero estimates. A group-selection property is also established, showing that highly correlated relevant regressors are retained together. Second, under exact collinearity, such as that caused by duplicated regressors, the observations determine only a parameter equivalence class. We characterize the selection induced by regularization within this class. Specifically, the first stage converges to the unique minimum-P-norm element of the minimum-L_1 set, while the second stage converges to the weighted L_1 solution set with finite-step recovery of the zero coordinates of the first-stage limit. Numerical examples and a robot-arm case study support the theory and illustrate the sparsity, stability, and robustness of the proposed method.**

***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, 62303452, 62588101, and T2293770, China Postdoctoral Science Foundation under Grant 2022M720159, RGC grant JLFS/P-501/24 for the CAS AMSS-PolyU Joint Laboratory in Applied Mathematics and Hong Kong Research Grant Council projects PolyU15300123,PolyU15300124. (Corresponding author: Ji-Feng Zhang.)

Jian Guo is with the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong (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. It has attracted considerable attention in a wide range of disciplines, including systems and control [1], signal processing [2], statistics [3], and machine learning [4]. 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.

Sparse identification has attracted considerable attention in statistics as well as in systems and control. In statistics, a wide range of sparsity-promoting techniques has been developed for variable selection by penalizing coefficient magnitudes. Representative examples include LASSO, elastic net [5], and adaptive LASSO [6]. Related developments include robust variable selection based on adaptive elastic-net S-estimators [7] and adaptive elastic-net estimation for sparse diffusion processes [8]. In the time series setting, [9] developed regularized methods for high-dimensional vector autoregressive (VAR) models, while [10] applied adaptive LASSO to large VAR models. More recently, [11] reviewed machine learning methods for time series forecasting. Although temporal dependence is explicitly considered in this literature, theoretical guarantees often rely on assumptions such as stationarity or 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 in [12] and the adaptive greedy algorithm in [13] for online sparse recovery, with applications to channel identification and linear prediction.

Despite their effectiveness in specific scenarios, these statistical and control approaches often rely on restrictive assumptions, such as independent and identically distributed (i.i.d.) stochastic inputs, known innovation distributions, or specific dependence structures, and their theoretical guarantees do not directly extend to more general stochastic settings. Related coherence-based sparse recovery theory provides finite-sample guarantees for the static model $b = Ax_0 + w$ with a fixed sensing matrix A [14]. More recently, [15] proposed coordinate-transformation and input-design methods to reduce mutual coherence in sparse system identification. In dynamic systems, however, the regression matrix A_n evolves with the sample size and may depend on past disturbances through feedback,

limiting the direct application of fixed-design results. These limitations are particularly relevant in applications such as control systems [16], financial time series [17], federated learning [18], and wireless communication channels [19], where independence, stationarity, or fixed-design assumptions may not hold. In closed-loop control, for example, inputs depend on past outputs and hence on past disturbances, yielding dependent and potentially non-stationary regressors. These issues motivate sparse identification methods with theoretical guarantees under stochastic, time-varying, and feedback-dependent settings.

Some progress has been made on sparse identification in the non-i.i.d. case. For example, [20] proposed a least-squares (LS) algorithm with weighted L_1 regularization for general observation sequences and established its convergence; [21] extended this framework to multivariate ARMA systems with exogenous inputs. More recently, [22] introduced an L_γ regularization method ($0 < \gamma < 1$) for sparse parameter identification in stochastic systems, proving exact support recovery and consistency of nonzero parameter estimates under general conditions. Despite these advances, a limitation of L_1-type regularization is that it can become unstable in high-dimensional settings, especially with highly correlated predictors [3], [5]. In practice, high dimensionality often leads to large sample correlations, making collinearity unavoidable. In system identification—an inverse problem—collinearity can cause numerical instability [23]. Moreover, estimating continuous-time impulse responses is ill-posed: small observation errors can cause 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 [24].

To address ill-conditioning, classical works by [25] proposed regularization methods. Later, [26] introduced kernel-based regularization methods in reproducing kernel Hilbert spaces, with further progress in kernel design [27] and in hyperparameter estimation [28]. 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 with a quadratic penalty 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 P-penalty. The weighted L_1 term promotes sparsity and support recovery, while the quadratic term stabilizes estimation under collinearity.
- For the increasingly severe collinearity case, where the condition number of the information matrix may diverge while the original parameter remains identifiable, this paper establishes global guarantees under non-i.i.d. and non-stationary observations via martingale techniques. Specifically, all zero coefficients are identified with probability one after finitely many observations, the nonzero coefficient estimates converge almost surely, and the estimators are asymptotically normal. Moreover, a grouping-effect bound is established for the proposed estimator, showing joint retention of highly correlated relevant regressors.
- For the exact-collinearity case, such as that caused by duplicated regressors, the original parameter is no longer uniquely identifiable and the observations determine only a parameter equivalence class. This paper characterizes the representative selected by regularization. The first-stage estimate converges to the unique minimum-P-norm element of the minimum-L_1-norm identifiable set, while the second-stage estimate converges to the corresponding weighted L_1 solution set with finite-step recovery of the zero coordinates. A set-valued view further provides a stability interpretation of the algorithm through the continuity of the minimum-P-norm selection.

The remainder of the paper is organized as follows. Notation is given at the end of Section I. Section II gives the problem formulation and the proposed two-stage regularized algorithm. Section III develops the theory for the non-exact collinearity case, including support recovery, convergence, asymptotic normality, and the group effect. Section IV addresses exact collinearity and establishes convergence for the corresponding regularized solution. Section V gives numerical and real-data examples, and Section VI concludes the paper.

Notation: Let $(\Omega, \mathcal{F}, \mathrm{P})$ be the probability space, $\omega \in \Omega$ a sample point, and $E[\cdot]$ the expectation operator. The symbols $\|\cdot\|_F$ and $\|\cdot\|$ denote the Frobenius norm and the Euclidean norm, respectively. Let $\mathbb{R}$ and $\mathbb{N}_+$ denote the sets of real numbers and positive integers. For $p \in \mathbb{N}_+$, I_p is the $p \times p$ identity matrix, $1_p = (1, \ldots, 1)^\top$, and $0_p = (0, \ldots, 0)^\top$. Let $\mathbb{S}^p_{++} := \{A \in \mathbb{R}^{p\times p} : A = A^\top,\ A \succ 0\}$. For $x \in \mathbb{R}^p$, $x(j)$ denotes its jth component, and $\mathrm{vec}(x(j))|_{j=1}^q = (x(1), \ldots, x(q))^\top$. The sign function is defined by $\mathrm{sign}(x) = 1$ if $x \geq 0$, and $\mathrm{sign}(x) = -1$ otherwise. For $A = A^\top \succeq 0$, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote its largest and smallest eigenvalues, $A^\dagger$ denotes its Moore–Penrose pseudoinverse, $\lambda^+_{\min}(A)$ denotes its smallest positive eigenvalue, and $\mathrm{Range}(A)$ denotes its range space. For a linear subspace $\mathcal{N} \subseteq \mathbb{R}^p$, $\mathcal{N}^\perp$ denotes its orthogonal complement, and $\Pi_{\mathcal{N}}$ denotes the Euclidean orthogonal projection onto $\mathcal{N}$. For positive sequences $\{a_k\}$ and $\{b_k\}$, $a_k = O(b_k)$ means that $a_k \leq cb_k$ eventually for some $c > 0$; $a_k = o(b_k)$ means that $a_k/b_k \to 0$; and $a_k \asymp b_k$ means that $c_1 b_k \leq a_k \leq c_2 b_k$ eventually for some $c_1, c_2 > 0$. For random sequences $\{x_k\}$ and $\{y_k\}$, $x_k = O_p(y_k)$ means that for any $\epsilon > 0$, there exist $M, N > 0$ such that $\mathrm{P}\{|x_k| \geq M|y_k|\} < \epsilon$ for all $k \geq N$; and $x_k = o_p(y_k)$ means that $x_k/y_k \xrightarrow{P} 0$, where $\xrightarrow{P}$ denotes 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 $A^* = \{j : \theta(j) = 0,\ j \in \{1, \dots, p\}\}$. Suppose that there are q nonzero elements in the vector θ. We then outline the following assumptions used in the theoretical analysis.

Assumptions. Denote the family of the σ-algebras $\mathcal{F}_k = \sigma\{y_k, \dots, y_1, u_k, \dots, u_0, w_k, \dots, w_1, w'_k, \dots, w'_1\},\ k \geq 1$, 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 as $n \to \infty$ a.s.

$$(a)\ \frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)} \to 0,\ (b)\ \frac{\lambda_{\max}(n)}{\lambda_{\min}(n)} \sqrt{\frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)}} \to 0.$$

Remark 1: Assumptions (A1) and (A2) allow the regressor sequence φ_k to be non-stationary and non-i.i.d. In (A1), $\{w_k, \mathcal{F}_k\}$ is a martingale difference sequence, which is more general than independent random variables and imposes weaker restrictions on temporal dependence. This permits w_{k+1} to depend on $\mathcal{F}_k$ and includes common distributions such as Gaussian. Assumption (A2) requires φ_k to be adapted to $\{\mathcal{F}_k\}$, as in PID control, adaptive regulation, and model reference control. Assumption (A3) characterizes the information growth of the regressors. Condition (a) is a classical sufficient condition for strong consistency of LS [29], while condition (b) is used for support recovery. It does not directly require system stability. For example, consider the open-loop FIR system $y_{k+1} = b_1 u_{1k} + b_2 u_{2k} + w_{k+1}$ with $u_{1k} = \xi_k$ and $u_{2k} = \xi_k + k^{-1/10}\eta_k$, where $\{\xi_k\}$ and $\{\eta_k\}$ are independent standard Gaussian sequences, independent of the disturbances. Then $\lambda_{\min}(n) \asymp n^{4/5}$ and $\lambda_{\max}(n) \asymp n$ almost surely. Thus, (A3) holds despite a diverging condition number, and this guarantee does not depend on the unknown coefficients. As an unstable example, consider the system $y_{k+1} = a_1 y_k + a_2 y_{k-1} + w_{k+1}$ with zero initial conditions and i.i.d. zero-mean Gaussian disturbances of positive variance. When the true coefficients are $(a_1, a_2) = (0, -\rho^2)$, $\rho > 1$, the system is unstable, but $\lambda_{\min}(n) \asymp \lambda_{\max}(n) \asymp \rho^{2n}$ almost surely, so (A3) also holds. These examples provide concrete sufficient settings.

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 [30] and large-scale system identification [31]. 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.

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 nearly linearly dependent, so $\sum_{k=1}^{n} \varphi_k \varphi_k^\top$ becomes ill-conditioned. This may make the estimates unstable and make variable selection more difficult. 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 $\theta_{n+1} = [\theta_{n+1}(1), \dots, \theta_{n+1}(p)]^\top$ 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, [20] 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 for $j = 1, \dots, p$,

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

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 [5], and it often selects only a single variable from a set of highly correlated regressors.

Regularized LS with quadratic penalties. To improve numerical stability in ill-conditioned problems, [23], [24] introduced a quadratic regularization term $\beta^\top P_n \beta$ with $P_n > 0$:

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

where $\bar{\lambda}_n > 0$ balances the data-fitting term and the quadratic penalty. This quadratic term improves stability under collinearity [23], but it does not directly induce sparsity.

Identification objective. Our goal is to construct a unified estimation framework for stochastic sparse systems under

non-i.i.d., non-stationary, and ill-conditioned settings. The framework aims to consistently recover the zero-coefficient set A^*, estimate the nonzero parameters, and improve prediction performance, with theoretical guarantees on parameter convergence, set consistency, and asymptotic normality.

It is worth noting that this paper considers two types of ill-conditioning. The first is non-exact collinearity under nonstationary and non-i.i.d. observations, where the information matrix remains nonsingular but may be severely ill-conditioned. In this case, θ is identifiable, and Section III establishes finite-step recovery of A^*. The second is persistent exact collinearity, which may make the information matrix rank deficient and destroy uniqueness of the identifiable parameter. This case is treated separately in Section IV, where the target is a regularized sparse representative of the corresponding equivalence class.

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). Inspired by the stabilizing role of quadratic regularization [23], [24], we first consider 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 [32], we adaptively reweight the L_1 term using the Stage-1 estimate (6), in the spirit of [20]. 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:

$$\hat\beta_n^0(j)=\beta_n^{(0)}(j)+\operatorname{sign}(\beta_n^{(0)}(j))\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}}+\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\right). \tag{7}$$

The refined problem is then

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

yielding the final estimator

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

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:

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

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

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

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

Remark 3: Assumption (A4) specifies the regularization rates relative to the information growth in (A3). The weighted L_1 penalty promotes sparse selection, while the positive definite quadratic penalty ensures strong convexity. Excessively large regularization parameters may cause regularization bias to dominate, whereas excessively small parameters may weaken sparse selection and finite-sample stabilization. The conditions in (A4) balance the two penalties against information growth to enable sparse selection while ensuring that the regularization bias vanishes asymptotically. Under (A3), one choice satisfying (A4), for sufficiently large n, is $\lambda_{1n}^{(1)}=\sqrt{\lambda_{\min}(n)\log\lambda_{\max}(n)}$, $\lambda_{1n}^{(2)}=\lambda_{\max}(n)\sqrt{\log\lambda_{\max}(n)/\lambda_{\min}(n)}$, and $\lambda_{2n}=\sqrt{\lambda_{\min}(n)}/\lambda_{\max}(P_n)$. These parameters can be computed from the observed regressors and the prescribed P_n, without knowledge of the true parameter vector. Thus, (A4) provides feasible tuning rules based on information growth rather than an additional structural restriction on the unknown system.

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 (10) and (12) 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 (10) and (12), the procedure in (13) reduces to the algorithm of [20].

Remark 5: We introduce a general quadratic regularization matrix P_n (rather than I_p) because a structured P_n can encode 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 [33], Laplacian-based regularization [34], and covariance–shrinkage forms [35]. Statistically, such P_n helps control variance and improves the bias–variance trade-off under collinearity [36]; from an optimization viewpoint, any $P_n>0$ maintains strict convexity and induces a better-conditioned proximal geometry [37]; numerically, it improves the conditioning of the regularized information

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 Remark 3.

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

$$\beta_n^{(0)} = \underset{\beta}{\operatorname{argmin}}\{\sum_{k=1}^{n}\left(y_{k+1}-\beta^\top\varphi_k\right)^2+\lambda_{1n}^{(1)}\sum_{j=1}^{p}|\beta(j)| + \lambda_{2n}\beta^\top P_n\beta\}. \tag{10}$$

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

$$\hat{\beta}_n^{(0)}(j) = \beta_n^{(0)}(j)+\operatorname{sign}(\beta_n^{(0)}(j))\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}} + \frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\right). \tag{11}$$

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

$$\hat{J}_n(\beta) = \sum_{k=1}^{n}\left(y_{k+1}-\beta^\top\varphi_k\right)^2+\lambda_{1n}^{(2)}\sum_{j=1}^{p}\frac{1}{|\hat{\beta}_n^{(0)}(j)|}|\beta(j)| + \lambda_{2n}\beta^\top P_n\beta \tag{12}$$

and obtain

$$\begin{cases}\widehat{\beta}_n = \left[\widehat{\beta}_n(1),\ldots,\widehat{\beta}_n(p)\right]^\top = \underset{\beta}{\operatorname{argmin}}\ \hat{J}_n(\beta),\\ \widehat{A}_n^* = \left\{j=1,\ldots,p\mid \widehat{\beta}_n(j)=0\right\}.\end{cases} \tag{13}$$

sample matrix and stabilizes estimation. These choices do not alter the asymptotic properties established in Section III-A.

Remark 6: The quadratic term $\lambda_{2n}\beta^\top P_n\beta$ may introduce a finite-sample shrinkage effect. Under Assumption (A4), this effect vanishes asymptotically and hence does not affect the convergence results in Subsection III-A. When $\sum_{k=1}^{n}\varphi_k\varphi_k^\top$ is nonsingular, one may also use the bias-reduced form $\widehat{\widehat{\beta}}_n = \left(I_p+\lambda_{2n}\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\right)^{-1}P_n\right)\widehat{\beta}_n$. This adjustment preserves the stated asymptotic properties.

Remark 7: The set $\widehat{A}_n^*$ in (13) is used to identify zero coefficients. In numerical implementation, due to finite precision and optimization tolerances, the estimated coefficients may not be exactly zero. Thus, one may use a threshold $\epsilon>0$, for example $\epsilon=10^{-10}$, and set $\widehat{A}_n^* = \left\{j\in\{1,\ldots,p\}:\ |\widehat{\beta}_n(j)|<\epsilon\right\}$. This gives a practical numerical version of the selection set.

III. Sparse Identification under Non-Exact Collinearity

This section studies the non-exact collinearity case, where the information matrix may be severely ill-conditioned but still contains sufficient excitation to identify the true parameter vector θ. We establish the large-sample properties of Algorithm 1 under non-i.i.d. and non-stationary observations, allowing the regressors to be feedback-dependent. Subsection III-A proves parameter convergence and finite-step identification of the zero coefficients. Subsection III-B establishes asymptotic normality of the nonzero estimates, and Subsection III-C studies the group-selection property under highly correlated regressors.

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., as $n\to\infty$, $\hat{\beta}_n(i)\to\theta(i),\ i=1,\ldots,p$ a.s.

For the set convergence analysis, we relabel the coordinates, if necessary, so that $\theta=(\theta_{10}^\top,0_{p-q}^\top)^\top,\ \theta_{10}\in\mathbb{R}^q$, where the first q components are nonzero. This relabeling is only for notational convenience and does not restrict the model.

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 parameter convergence, while Theorem 2 establishes finite-step identification of the zero components of θ. We now prove these results. To this end, we first recall two classical results used in the subsequent analysis.

Lemma 1: [38] Consider system (1). If Assumptions (A1) and (A2) hold, then $\left\|\left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top\right)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k w_{k+1}\right\| = O\left(\sqrt{\log\lambda_{\max}(n)}\right)$, a.s.

Lemma 2: [38] Consider system (1). If Assumptions (A1) and (A2) hold, then the estimation error of the LS algorithm in (2) satisfies $\|\theta_{n+1}-\theta\| = O\left(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\right)$, 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:

$$J_n(\beta,\alpha_{1n},\alpha_{2n},\eta_n) = \sum_{k=1}^{n}\left(y_{k+1}-\beta^\top\varphi_k\right)^2+\alpha_{2n}\beta^\top P_n\beta + \alpha_{1n}\sum_{j=1}^{p}\eta_n(j)|\beta(j)|, \tag{14}$$

$$\beta_n(\alpha_{1n},\alpha_{2n},\eta_n) = \underset{\beta}{\operatorname{argmin}} J_n(\beta,\alpha_{1n},\alpha_{2n},\eta_n). \tag{15}$$

Lemma 3: Under Assumptions (A1) and (A2), the auxiliary estimate $\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)$ defined in (15) admits the

following non-asymptotic bound:

$$\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\| = O\Bigg(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}} + \frac{\alpha_{1n}\sqrt{\sum_{j=1}^{q}(\eta_n(j))^2}+\alpha_{2n}\lambda_{\max}(P_n)\|\theta\|}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\Bigg). \quad (16)$$

Proof: The proof is given in Appendix. ■

Lemma 3 yields a non-asymptotic error bound for the estimate in (11), which implies the convergence of the Stage 1 estimate and supports the convergence analysis of Algorithm 1.

Corollary 1: Under Assumptions (A1) and (A2), the initial estimate $\hat{\beta}_n^{(0)}$ defined in (11) satisfies

$$\|\hat{\beta}_n^{(0)}-\theta\| = O\Bigg(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}} + \frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\|\theta\|}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\Bigg). \quad (17)$$

Proof: From (5), (6), (14), and (15), we have $\beta_n^{(0)} = \beta_n(\lambda_{1n}^{(1)},\lambda_{2n},1_p)$. The bound (17) follows directly from Lemma 3 together with (11). ■

Remark 8: The explicit parameter-norm term in (17) arises from quadratic regularization. The quadratic penalty increases the lower eigenvalue bound through $\lambda_{2n}\lambda_{\min}(P_n)$, while introducing a regularization bias that is linear in $\|\theta\|$. Under (A4)(a), the coefficient of $\|\theta\|$ tends to zero. Thus, for each fixed θ, this contribution vanishes and does not prevent consistency under (A3)(a) and the corresponding tuning conditions, which do not require knowledge of $\|\theta\|$.

We now present the proofs of Theorems 1 and 2.

Proof of Theorem 1. From (8) and (15), define $\hat{\eta}_n(i) := 1/|\hat{\beta}_n^{(0)}(i)|$, $i=1,\dots,p$, and $\hat{\eta}_n := (\hat{\eta}_n(1),\dots,\hat{\eta}_n(p))^\top$. 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 $i=1,\dots,q$, 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 $\sum_{j=1}^{q}\hat{\eta}_n(j)^2 = O(1)$, and by Lemma 3, it follows $\|\hat{\beta}_n-\theta\| = \|\beta_n(\lambda_{1n}^{(2)},\lambda_{2n},\hat{\eta}_n)-\theta\| = O\Bigg(\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}+\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\Bigg)$. 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 $\hat{\beta}_{1n} = \theta_{10}+t_n u_{1n}$, $\hat{\beta}_{2n} = t_n u_{2n}$, where $u_{1n}\in\mathbb{R}^q$ and $u_{2n}\in\mathbb{R}^{p-q}$. In addition, denote

$$\sum_{k=1}^{n}\varphi_k\varphi_k^\top = \begin{bmatrix}\Phi_n^{(11)} & \Phi_n^{(12)}\\ \Phi_n^{(21)} & \Phi_n^{(22)}\end{bmatrix}, \quad \varphi_k = \begin{bmatrix}\varphi_k^{(1)}\\ \varphi_k^{(2)}\end{bmatrix}, \quad (18)$$

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 (12), it follows that

$$\hat{J}_n(\hat{\beta}_n)-\hat{J}_n(\hat{\beta}_{1n},0_{p-q})\le 0. \quad (19)$$

Noting (1), a direct calculation for (12) yields

$$\begin{aligned}\hat{J}_n(\hat{\beta}_n) &= \sum_{k=1}^{n}w_{k+1}^2 + t_n^2u_{1n}^\top\Phi_n^{(11)}u_{1n} + t_n^2u_{2n}^\top\Phi_n^{(22)}u_{2n}\\ &+2t_n^2\sum_{k=1}^{n}\big(\varphi_k^{(1)\top}u_{1n}\big)\big(\varphi_k^{(2)\top}u_{2n}\big) \quad (20)\\ &-2t_n\sum_{k=1}^{n}\varphi_k^{(1)\top}u_{1n}w_{k+1} - 2t_n\sum_{k=1}^{n}\varphi_k^{(2)\top}u_{2n}w_{k+1}\\ &+\lambda_{2n}t_n^2\big(u_{1n}^\top P_{n,11}u_{1n}+2u_{1n}^\top P_{n,12}u_{2n}+u_{2n}^\top P_{n,22}u_{2n}\big)\\ &+\lambda_{1n}^{(2)}\sum_{j=1}^{q}\frac{1}{|\hat{\beta}_n^{(0)}(j)|}\big|\hat{\beta}_{1n}(j)\big|+\lambda_{1n}^{(2)}t_n\sum_{j=1}^{p-q}\frac{|u_{2n}(j)|}{|\hat{\beta}_n^{(0)}(j+q)|},\end{aligned}$$

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,

$$\begin{aligned}&\hat{J}_n(\hat{\beta}_{1n},0_{p-q})\\ &=\sum_{k=1}^{n}w_{k+1}^2 + t_n^2u_{1n}^\top\Phi_n^{(11)}u_{1n} - 2t_n\sum_{k=1}^{n}\varphi_k^{(1)\top}u_{1n}w_{k+1}\\ &\quad+\lambda_{2n}t_n^2u_{1n}^\top P_{n,11}u_{1n}+\lambda_{1n}^{(2)}\sum_{j=1}^{q}\frac{1}{|\hat{\beta}_n^{(0)}(j)|}\big|\hat{\beta}_{1n}(j)\big|. \quad (21)\end{aligned}$$

Subtracting (21) from (20), we obtain

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

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

$$\begin{aligned}&T_n^{(1)}+T_n^{(2)}\\ &=2t_n^2\sum_{k=1}^{n}\big(\varphi_k^{(1)\top}u_{1n}\big)\big(\varphi_k^{(2)\top}u_{2n}\big)+t_n^2\sum_{k=1}^{n}\big(\varphi_k^{(2)\top}u_{2n}\big)^2\\ &\ge 2t_n^2\,u_{1n}^\top\Bigg(\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(2)\top}\Bigg)u_{2n}\\ &\ge -2t_n^2\|u_{1n}\|\,\|u_{2n}\|\,\Bigg\|\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(2)\top}\Bigg\|\\ &\ge -2c_2\,t_n^2\bar{U}\,\|u_{2n}\|\,\lambda_{\max}(n) \quad (23)\end{aligned}$$

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

$$\begin{aligned}T_n^{(3)} &= -2t_n\sum_{k=1}^{n}\varphi_k^{(2)\top}u_{2n}\,w_{k+1}\\ &\ge -2t_n\|u_{2n}\|\big\|\Phi_n^{(22)\frac{1}{2}}\big\|\,\Bigg\|\Big(\Phi_n^{(22)}\Big)^{-\frac{1}{2}}\sum_{k=1}^{n}\varphi_k^{(2)}w_{k+1}\Bigg\|\\ &\ge -2t_n\|u_{2n}\|\,\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n). \quad (24)\end{aligned}$$

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 $c_4>c_3\geq 1$ such that $c_3\gamma_n\leq|\hat{\beta}_n^{(0)}(j)|\leq c_4\gamma_n,\ j=q+1,\dots,p$, which implies

$$\frac{1}{c_4\gamma_n}\leq\frac{1}{|\hat{\beta}_n^{(0)}(j)|}\leq\frac{1}{c_3\gamma_n},\quad j=q+1,\dots,p. \tag{25}$$

Therefore, using (25) and the block partition of P_n, we obtain

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

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

$$T_n^{(4)}+T_n^{(5)}\geq-2\lambda_{2n}t_n^2\bar{U}\lambda_{\max}(P_n)\|u_{2n}\|+\frac{\lambda_{1n}^{(2)}t_n}{c_4\gamma_n}\|u_{2n}\| \tag{26}$$

Combining (22) and (26), we have the following overall lower bound for the terms:

$$\begin{aligned}&T_n^{(1)}+T_n^{(2)}+T_n^{(3)}+T_n^{(4)}+T_n^{(5)}\\&\geq\frac{\lambda_{1n}^{(2)}t_n}{c_4\gamma_n}\|u_{2n}\|\Bigg(1-2c_2c_4\bar{U}\,\frac{t_n\gamma_n\,\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}-2c_4\bar{U}t_n\gamma_n\,\lambda_{2n}\\&\quad\cdot\frac{\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}-2c_4\,\frac{\gamma_n\,\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\Bigg).\end{aligned} \tag{27}$$

Hence, by (19), (22), and (27), we have

$$\begin{aligned}0\geq&\|u_{2n}\|\Bigg(1-2c_2c_4\bar{U}\,\frac{t_n\gamma_n\,\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}-2c_4\bar{U}t_n\gamma_n\,\lambda_{2n}\\&\cdot\frac{\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}-2c_4\,\frac{\gamma_n\,\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\Bigg).\end{aligned} \tag{28}$$

We next estimate the three terms in (28).

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

$$\begin{aligned}&\frac{t_n\gamma_n\lambda_{\max}(n)}{\lambda_{1n}^{(2)}}\leq\frac{\log\lambda_{\max}(n)\lambda_{\max}(n)}{\lambda_{\min}(n)\lambda_{1n}^{(2)}}\\&+\frac{\sqrt{\log\lambda_{\max}(n)}\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)}{\lambda_{\min}(n)^{\frac{3}{2}}\lambda_{1n}^{(2)}}\\&+\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\\&+\frac{\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)}{\lambda_{\min}(n)^2}\frac{\lambda_{1n}^{(2)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\\&\xrightarrow[n\to\infty]{}0\text{ a.s.}\end{aligned} \tag{29}$$

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

$$\begin{aligned}&\frac{t_n\gamma_n\,\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\leq C\Bigg[\frac{\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\\&+\frac{\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\cdot\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{\min}(n)+\lambda_{2n}\lambda_{\min}(P_n)}\Bigg]\\&\xrightarrow[n\to\infty]{}0\text{ a.s. (by (A3)(a),(A4)(b) and (A4)(a),(A4)(b))}\end{aligned} \tag{30}$$

(iii) Third term. Finally,

$$\begin{aligned}&\frac{\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)\gamma_n}{\lambda_{1n}^{(2)}}\leq\frac{\log\lambda_{\max}(n)\lambda_{\max}(n)^{\frac{1}{2}}}{\lambda_{\min}(n)^{\frac{1}{2}}\lambda_{1n}^{(2)}}\\&+\frac{\big(\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)\big)\lambda_{\max}(n)^{\frac{1}{2}}\log^{\frac{1}{2}}\lambda_{\max}(n)}{\lambda_{\min}(n)\lambda_{1n}^{(2)}}\\&=\frac{\log\lambda_{\max}(n)\lambda_{\max}(n)}{\lambda_{\min}(n)\lambda_{1n}^{(2)}}\frac{\lambda_{\min}(n)^{\frac{1}{2}}}{\lambda_{\max}(n)^{\frac{1}{2}}}\\&+\frac{\lambda_{1n}^{(1)}+\lambda_{2n}\lambda_{\max}(P_n)}{\lambda_{1n}^{(2)}}\frac{\lambda_{\max}(n)}{\lambda_{\min}(n)}\sqrt{\frac{\log\lambda_{\max}(n)}{\lambda_{\min}(n)}}\frac{\lambda_{\min}(n)^{\frac{1}{2}}}{\lambda_{\max}(n)^{\frac{1}{2}}}\\&\xrightarrow[n\to\infty]{}0\text{ a.s. (by (A4)(c) and (A4)(d))}\end{aligned} \tag{31}$$

Combining (29), (30), and (31), we deduce that each of the three terms inside the parentheses of (28) 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 these estimates into (28) gives $\|u_{2n}\|=0$ for all sufficiently large n. Hence $\widehat{A}_n^*=A^*$ eventually a.s. 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

$$\begin{aligned}&R_n^{-1}\Phi_n^{(11)}\xrightarrow{P}I_q,\quad\max_{1\leq k\leq n}\|R_n^{-1/2}\varphi_k^{(1)}\|\xrightarrow{P}0,\\&\lim_{k\to\infty}E\left(w_{k+1}^2\mid\mathcal{F}_k\right)=\sigma^2\text{ a.s. for some constant }\sigma,\end{aligned} \tag{32}$$

where $\varphi_k^{(1)}$ and $\Phi_n^{(11)}$ are defined in (18). 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 $\|\upsilon_n\|\leq 1$, 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

$$s_n^{-1}\,\upsilon_n^\top\left(\left(I_q+\lambda_{2n}\,(\Phi_n^{(11)})^{-1}P_{n,11}\right)\hat{\beta}_{1n}-\theta_{10}\right)\xrightarrow{\mathrm{d}}N(0,1), \tag{33}$$

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

Remark 9: The deterministic matrix R_n in Theorem 3 is a stability condition for $\Phi_n^{(11)}$. This condition is essential for asymptotic normality; without it, asymptotic normality may fail, as shown by Example 3 in [39]. Moreover, when $\{\varphi_k^{(1)}\}$

is predetermined, one may take $R_n = \Phi_n^{(11)}$. If $\varphi_n^{(1)}\varphi_n^{(1)\top}$ is stationary and ergodic with positive definite covariance, one may take $R_n = n\,E\Big[\varphi_n^{(1)}\varphi_n^{(1)\top}\Big]$, as in [40].

Proof: Write $\hat{J}_n(\beta)$ in (12) 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 $|\hat{\beta}_{1n}(j)| \geq c$ for all $j = 1, \ldots, q$ and all $n \geq N(\omega)$. 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,

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

Using (1) and $\theta = [\theta_{10}^\top, 0_{p-q}^\top]^\top$, we have $y_{k+1} = \theta_{10}^\top\varphi_k^{(1)} + w_{k+1}$. Thus, it follows

$$
\begin{aligned}
&\sum_{k=1}^{n}\varphi_k^{(1)}\varphi_k^{(1)\top}\Big(\big(I_q + \lambda_{2n}(\Phi_n^{(11)})^{-1}P_{n,11}\big)\hat{\beta}_{1n} - \theta_{10}\Big) \\
&= -\sum_{k=1}^{n}\hat{\beta}_{2n}^\top\varphi_k^{(2)}\varphi_k^{(1)} + \sum_{k=1}^{n}\varphi_k^{(1)}w_{k+1} - \lambda_{2n}P_{n,12}\hat{\beta}_{2n} \\
&\quad - \frac{1}{2}\lambda_{1n}^{(2)}\,\mathrm{vec}\big(\mathrm{sign}(\hat{\beta}_{1n}(j))\,|\hat{\beta}_n^{(0)}(j)|^{-1}\big)\Big|_{j=1}^{q}, \qquad (34)
\end{aligned}
$$

Premultiplying (34) by $(\Phi_n^{(11)})^{-1}$ and then by $s_n^{-1}\upsilon_n^\top$ gives

$$
\begin{aligned}
&s_n^{-1}\upsilon_n^\top\Big(\big(I_q + \lambda_{2n}(\Phi_n^{(11)})^{-1}P_{n,11}\big)\hat{\beta}_{1n} - \theta_{10}\Big) \\
&= -s_n^{-1}\upsilon_n^\top(\Phi_n^{(11)})^{-1}\Big(\sum_{k=1}^{n}\hat{\beta}_{2n}^\top\varphi_k^{(2)}\varphi_k^{(1)} + \lambda_{2n}P_{n,12}\hat{\beta}_{2n}\Big) \\
&\quad + s_n^{-1}\sum_{k=1}^{n}\upsilon_n^\top(\Phi_n^{(11)})^{-1}\varphi_k^{(1)}w_{k+1} \qquad (35) \\
&\quad - \frac{1}{2}\lambda_{1n}^{(2)}\,s_n^{-1}\upsilon_n^\top(\Phi_n^{(11)})^{-1}\,\mathrm{vec}\big(\mathrm{sign}(\hat{\beta}_{1n}(j))\,|\hat{\beta}_n^{(0)}(j)|^{-1}\big)\Big|_{j=1}^{q}.
\end{aligned}
$$

For the first term on the right-hand side of (35), by Theorem 2 we have $\hat{\beta}_{2n} = 0$ eventually, a.s., hence

$$
\mathrm{P}\left(\lim_{n\to\infty}\sum_{k=1}^{n}\hat{\beta}_{2n}^\top\varphi_k^{(2)}\varphi_k^{(1)} + \lambda_{2n}P_{n,12}\hat{\beta}_{2n} = 0\right) = 1. \qquad (36)
$$

For the last term on the right-hand side of (35), 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 $|\hat{\beta}_n^{(0)}(j)| \geq c_5$ for $j = 1, \ldots, q$ when n is sufficiently large. Moreover, by the definition of s_n and the first condition in (32), 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

$$
\begin{aligned}
&\left|\lambda_{1n}^{(2)}s_n^{-1}\,\upsilon_n^\top\left(\Phi_n^{(11)}\right)^{-1}\,\mathrm{vec}\Big(\mathrm{sign}(\hat{\beta}_{1n}(j))|\hat{\beta}_n^{(0)}(j)|^{-1}\Big)\Big|_{j=1}^{q}\right| \\
&\leq \frac{\lambda_{1n}^{(2)}}{\lambda_{\min}(n)^{1/2}}\cdot\frac{s_n^{-1}}{\lambda_{\min}(n)^{1/2}}\cdot\frac{q^{1/2}}{c_5} \xrightarrow{P} 0 \qquad (37)
\end{aligned}
$$

Hence, combining (35), (36) and (37), we obtain

$$
\begin{aligned}
&s_n^{-1}\,\upsilon_n^\top\Big(\big(I_q + \lambda_{2n}(\Phi_n^{(11)})^{-1}P_{n,11}\big)\hat{\beta}_{1n} - \theta_{10}\Big) \\
&= s_n^{-1}\sum_{k=1}^{n}\upsilon_n^\top\left(\Phi_n^{(11)}\right)^{-1}\varphi_k^{(1)}w_{k+1} + o_p(1) \qquad (38)
\end{aligned}
$$

In view of (32) and (38), and by Slutsky's theorem [41], to prove (33) it suffices to show that $s_n^{-1}\sum_{k=1}^{n}\upsilon_n^\top R_n^{-1}\varphi_k^{(1)}w_{k+1} \xrightarrow{\mathrm{d}} N(0,1)$. Along the lines of [39], the desired limit follows from the martingale central limit theorem of [42]. ∎

Remark 10: Theorem 3 further establishes asymptotic normality for the estimator proposed in [20]. Denote this estimator by $\bar{\beta}_n = (\bar{\beta}_{1n}^\top, \bar{\beta}_{2n}^\top)^\top$ with $\bar{\beta}_{1n} \in \mathbb{R}^q$. If $\lambda_{2n} = 0$, $\lambda_{1n}^{(1)} = 0$, and $\lambda_{1n}^{(2)}/\lambda_{\min}^{1/2}(n) \xrightarrow{P} 0$, with $\lambda_{1n}^{(2)}$ satisfying Assumption (A4), then $s_n^{-1}\upsilon_n^\top(\bar{\beta}_{1n} - \theta_{10}) \xrightarrow{\mathrm{d}} N(0,1)$.

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 [43]: $y_m = \sum_{d=1}^{D}A_d x_{d,m} + n_m$, 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 most are zero. Hence the channel vectors are sparse. Moreover, these vectors are naturally grouped by the user index d, and variables within the same group are often highly correlated; for example, the input matrix A_d may be low-rank. Thus, an effective channel estimation method should both remove insignificant paths and retain correlated paths within the same group. Algorithm 1 achieves these two goals by combining sparse selection with the group effect. Theorem 4 further shows that Algorithm 1 has this group-selection property.

It is worth noting that a widely used method for enforcing group sparsity is the Group Lasso [44], 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

$$
\begin{aligned}
\big|(P_n\hat{\beta}_n)(i) - (P_n\hat{\beta}_n)(j)\big| &\leq \sqrt{\sum_{k=1}^{n}y_{k+1}^2}\,\sqrt{\sum_{k=1}^{n}\big(\varphi_k(i) - \varphi_k(j)\big)^2} \\
&\quad\times\left(\frac{1}{\lambda_{2n}} + \frac{\lambda_{1n}^{(2)}}{2\lambda_{2n}^2\,\min\big(|\hat{\beta}_n^{(0)}(i)|,\,|\hat{\beta}_n^{(0)}(j)|\big)^2}\right), \qquad (39)
\end{aligned}
$$

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

Remark 11: Inequality (39) shows that the P_n-weighted coefficients $(P_n\hat{\beta}_n)(i)$ and $(P_n\hat{\beta}_n)(j)$ are close whenever $\sum_{k=1}^n \big(\varphi_k(i)-\varphi_k(j)\big)^2$ is small. After standardization, this corresponds to the case where the two regressors are highly correlated. Hence the quadratic P_n-term promotes a group effect by coupling the estimates of highly correlated variables. Both the quadratic term and the weighted L_1 term contribute to this property. In the pure L_1 case ($\lambda_{2n}=0$), this coupling mechanism disappears, and (39) no longer provides a meaningful group-effect bound.

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

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

Hence,

$$\lambda_{2n}\Big|(P_n\hat{\beta}_n)(i)-(P_n\hat{\beta}_n)(j)\Big|\le\frac{\lambda_{1n}^{(2)}}{2}\Big||\hat{\beta}_n^{(0)}(i)|^{-1}-|\hat{\beta}_n^{(0)}(j)|^{-1}\Big|+\left|\sum_{k=1}^n\big(y_{k+1}-\hat{\beta}_n^\top\varphi_k\big)\big(\varphi_k(i)-\varphi_k(j)\big)\right|. \tag{40}$$

Since $\hat{\beta}_n$ is a minimizer, $\hat{J}_n(0_p)\ge\hat{J}_n(\hat{\beta}_n)$ implies $\sum_{k=1}^n\big(y_{k+1}-\hat{\beta}_n^\top\varphi_k\big)^2\le\sum_{k=1}^n y_{k+1}^2$. Therefore, by Cauchy–Schwarz inequality,

$$\left|\sum_{k=1}^n\big(y_{k+1}-\hat{\beta}_n^\top\varphi_k\big)\big(\varphi_k(i)-\varphi_k(j)\big)\right|\le\sqrt{\sum_{k=1}^n y_{k+1}^2}\sqrt{\sum_{k=1}^n\big(\varphi_k(i)-\varphi_k(j)\big)^2}. \tag{41}$$

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

$$\Big||\hat{\beta}_n^{(0)}(i)|^{-1}-|\hat{\beta}_n^{(0)}(j)|^{-1}\Big|=\frac{\big||\hat{\beta}_n^{(0)}(i)|-|\hat{\beta}_n^{(0)}(j)|\big|}{\xi^2}\le\frac{|\hat{\beta}_n^{(0)}(i)-\hat{\beta}_n^{(0)}(j)|}{\big(\min\{|\hat{\beta}_n^{(0)}(i)|,|\hat{\beta}_n^{(0)}(j)|\}\big)^2}. \tag{42}$$

Plugging (41) and (42) into (40) gives

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

This completes the proof. ∎

IV. Exact Collinearity and Regularized Sparse Representatives

In this section, we study the behavior of Algorithm 1 for system (1) under exact collinearity, where a nontrivial linear dependence among the regressor components persists throughout the observation sequence. In this setting, the LS criterion has nonunique minimizers, and ℓ_1 regularization does not generally ensure uniqueness. For example, proportional normalized design columns yield unit mutual coherence, under which the sufficient condition for the coherence-based uniqueness guarantee for ℓ_1-regularized LS cannot hold for any nonzero sparsity level, see [14, Theorem 3]. We therefore examine what can be identified from the observations before studying the sparse representatives selected by regularization.

A. Sparse Identifiability under Exact Collinearity

The observations provide information only along the directions spanned by the regressors. Since the regressors and hence G_n are random, we first characterize these directions along individual sample paths. The following lemma shows that the corresponding subspaces eventually stabilize.

Lemma 4: Let $\mathcal{S}_\infty(\omega)=\operatorname{span}\{\varphi_k(\omega):k\ge1\}$. For $\mathbb{P}$-a.e. ω, there exists a finite integer $n_0(\omega)\ge1$ such that, for all $n\ge n_0(\omega)$, $\operatorname{range}(G_n(\omega))=\mathcal{S}_\infty(\omega)$, $\ker(G_n(\omega))=\mathcal{S}_\infty(\omega)^\perp$.

Proof: For each $\omega\in\Omega$, subspaces $\{\operatorname{range}(G_n(\omega))\}_{n\ge1}$ are nested and have dimensions bounded by p. Hence, they eventually stabilize, and their eventual value is $\mathcal{S}_\infty(\omega)$. The kernel identity follows from the symmetry of $G_n(\omega)$. ∎

Although stabilization holds pathwise, the limiting subspace depend on the sample path. We consider exact collinearity with a common, deterministic limiting subspace, as below.

Assumption 1 (Exact collinearity): There exists a deterministic linear subspace $\mathcal{S}\subsetneq\mathbb{R}^p$ such that $\mathcal{S}_\infty(\omega)=\mathcal{S}$, a.s.

Set $\mathcal{N}=\mathcal{S}^\perp$. Lemma 4 gives $\ker(G_n(\omega))=\mathcal{N}$, $n\ge n_0(\omega)$, a.s. and the unidentifiable directions are independent of the sample path. Parameters differing by a vector in $\mathcal{N}$ produce identical noiseless predictions a.s., which motivates the following definition.

Definition 1 (Identifiable equivalence class): Under Assumption 1, the identifiable equivalence class of θ is defined by $[\theta]=\theta+\mathcal{N}$.

Remark 12: Assumption 1 describes the exact collinearity structure. The subspace $\mathcal{N}$ consists of the directions that are never excited by the regressors, while $\mathcal{N}^\perp$ contains the identifiable directions. Indeed, for any $v\in\mathcal{N}$, by $v\in\ker(G_n)$, $\forall n\ge n_0$, we have $0=v^\top G_n v=\sum_{k=1}^n(v^\top\varphi_k)^2$, $\forall n\ge n_0$. It follows that $v^\top\varphi_k=0$, $\forall k\ge1$. Therefore, $G_nv=0$, $\forall n\ge1$. Thus every direction in $\mathcal{N}$ is exactly unobservable from the regressor sequence.

Moreover, for any $\beta\in[\theta]$, one has $\beta-\theta\in\mathcal{N}$. Hence $\beta^\top\varphi_k=\theta^\top\varphi_k$, $\forall k\ge1$. Therefore, for system (1), $y_{k+1}=\theta^\top\varphi_k+w_{k+1}=\beta^\top\varphi_k+w_{k+1}$, $\forall k\ge1$. This shows that every parameter in $[\theta]$ generates exactly the same noiseless prediction sequence as the true parameter θ. Hence, under exact collinearity, the data cannot distinguish θ from any other element of $[\theta]$. The estimation target should therefore be a stable representative from $[\theta]$, rather than θ itself.

When sparsity is further imposed, the selection within $[\theta]$ should favor representatives with small support. We therefore introduce the following set.

Definition 2 (Sparse identifiable set): Define the sparse identifiable set $\mathcal{S}_0=\{\beta\in[\theta]:\|\beta\|_0=s_0\}$ with $s_0=\min_{\beta\in[\theta]}\|\beta\|_0$.

The set $\mathcal{S}_0$ consists of the sparsest representatives in the identifiable equivalence class $[\theta]$. Since $\theta\in[\theta]$, it is nonempty

and satisfies $s_0 \le \|\theta\|_0$. Thus, under exact collinearity, the result characterizes regularization-selected representatives and does not, without additional assumptions, guarantee recovery of an ℓ_0-sparsest representative.

B. Regularized Sparse Algorithm and Its Convergence

The ordinary LS problem is $\min_{\beta\in\mathbb{R}^p} L_n(\beta) := \sum_{k=1}^n \left(y_{k+1} - \beta^\top \varphi_k\right)^2$. Since G_n is singular under exact collinearity, the LS solution is generally not unique. The same issue also occurs for the L_1-regularized criterion. Under exact collinearity, the ℓ_1-regularized criterion need not have a unique minimizer. For proportional design columns, the coherence-based uniqueness condition in [14] fail.

We therefore use Algorithm 1, where the additional quadratic term is used to select a unique representative. Under Assumption 1, the identifiable equivalence class is $[\theta] = \theta + \mathcal{N}$. We define the minimum-$L_1$-norm representative set by

$$\mathcal{S}_1 = \arg\min_{\beta\in[\theta]} \|\beta\|_1. \tag{44}$$

Since $[\theta]$ is a nonempty closed affine set and $\|\cdot\|_1$ is coercive and convex, $\mathcal{S}_1$ is nonempty, compact, and convex. Let $P \in \mathbb{S}^p_{++}$. Define

$$\beta^\star_{1,P} = \arg\min_{\beta\in\mathcal{S}_1} \beta^\top P \beta. \tag{45}$$

Since $P \in \mathbb{S}^p_{++}$, the above minimizer is unique. For $n \ge n_0$, define $\lambda^+_{\min}(n) = \inf_{u\in\mathcal{N}^\perp,\ \|u\|=1} u^\top G_n u$. Under Assumption 1, G_n is positive definite on $\mathcal{N}^\perp$, and hence $\lambda^+_{\min}(n) > 0$.

Consider the regularized sparse estimator (10) with $P_n \equiv P$. Since $P \in \mathbb{S}^p_{++}$ and $\lambda_{2n} > 0$, the objective function is strongly convex. Hence $\beta^{(0)}_n$ is unique for each n. Let $M_n = \sum_{k=1}^n \varphi_k w_{k+1}$. The following lemma shows that, even under exact collinearity, the first-stage regularized estimator $\beta^{(0)}_n$ in Algorithm 1 is well defined and converges to the P-minimum representative in the L_1-minimum identifiable set, with explicit rates for both identifiable and unidentifiable components.

Lemma 5: Set $e_n = \sqrt{\frac{\log\lambda_{\max}(n)}{\lambda^+_{\min}(n)} + \frac{\lambda^{(1)}_{1n} + \lambda_{2n}\lambda_{\max}(P)}{\lambda^+_{\min}(n) + \lambda_{2n}\lambda_{\min}(P)}}$ and $h_n = e_n + \frac{\log\lambda_{\max}(n) + \lambda_{2n}}{\lambda^{(1)}_{1n}}$. Suppose that $\lambda_{\max}(n) \to \infty$ and $h_n \to 0$, a.s.. Under Assumptions (A1)–(A2) and Assumption 1, $\operatorname{dist}(\beta^{(0)}_n, [\theta]) = O(e_n)$, a.s., and $\operatorname{dist}(\beta^{(0)}_n, \mathcal{S}_1) = O(h_n)$, a.s. Moreover, $\|\beta^{(0)}_n - \beta^\star_{1,P}\| = O\left(e_n + \mathbf{1}_{\{\mathcal{N}\neq\{0\}\}}\sqrt{h_n + \frac{\log\lambda_{\max}(n)}{\lambda_{2n}} + \frac{\lambda^{(1)}_{1n}}{\lambda_{2n}} e_n}\right)$, a.s.

Proof: Let $d_n = \beta^{(0)}_n - \beta^\star_{1,P}$. Since $\beta^\star_{1,P} \in [\theta] = \theta + \mathcal{N}$, by Remark 12, we have $y_{k+1} - (\beta^\star_{1,P})^\top \varphi_k = w_{k+1},\ k \ge 1$.

Let $U \in \mathbb{R}^{p\times r}$ be an orthonormal basis matrix of $\mathcal{N}^\perp$. Then $\Pi_{\mathcal{N}^\perp} x = UU^\top x$ and $\|\Pi_{\mathcal{N}^\perp} x\| = \|U^\top x\|$. Since $\ker(G_n) = \mathcal{N}$ for all sufficiently large n, the matrix $\bar{G}_n = U^\top G_n U$ is positive definite on $\mathbb{R}^r$, and $\lambda_{\min}(\bar{G}_n) = \lambda^+_{\min}(n)$, $\lambda_{\max}(\bar{G}_n) \le \lambda_{\max}(n)$. Moreover, for any $\beta \in [\theta] = \theta + \mathcal{N}$, one has $U^\top\beta = U^\top\theta$. In particular, $U^\top\beta^\star_{1,P} = U^\top\theta$. Therefore, on the identifiable subspace $\mathcal{N}^\perp$, the regression model reduces to $y_{k+1} = \bar{\varphi}_k^\top U^\top\theta + w_{k+1}$ with $\bar{\varphi}_k = U^\top\varphi_k$.

By the optimality of $\beta^{(0)}_n$,

$$0 \ge \sum_{k=1}^n \left(y_{k+1} - (\beta^{(0)}_n)^\top\varphi_k\right)^2 - \sum_{k=1}^n \left(y_{k+1} - (\beta^\star_{1,P})^\top\varphi_k\right)^2 + \lambda^{(1)}_{1n}\left(\|\beta^{(0)}_n\|_1 - \|\beta^\star_{1,P}\|_1\right) + \lambda_{2n}\left((\beta^{(0)}_n)^\top P\beta^{(0)}_n - (\beta^\star_{1,P})^\top P\beta^\star_{1,P}\right).$$

Using the system equation and $G_n(\beta^\star_{1,P} - \theta) = 0$, we obtain

$$d_n^\top G_n d_n + \lambda^{(1)}_{1n}\left(\|\beta^{(0)}_n\|_1 - \|\beta^\star_{1,P}\|_1\right) + \lambda_{2n}\left((\beta^{(0)}_n)^\top P\beta^{(0)}_n - (\beta^\star_{1,P})^\top P\beta^\star_{1,P}\right) \le 2d_n^\top M_n. \tag{46}$$

Applying the same transformation to the identifiable subspace $\mathcal{N}^\perp$ as above, and then using Lemma 1, we can obtain $\|G_n^{\dagger/2} M_n\| = O\left(\sqrt{\log\lambda_{\max}(n)}\right)$, a.s. Set $Z_n := \|G_n^{\dagger/2} M_n\|^2$. By (46), $\|\beta^{(0)}_n\|_1 \le \|\beta^\star_{1,P}\|_1 + (\lambda_{2n}(\beta^\star_{1,P})^\top P\beta^\star_{1,P} + Z_n)/\lambda^{(1)}_{1n} = O(1)$, a.s. The optimality condition gives $G_n d_n = M_n - \frac{\lambda^{(1)}_{1n}}{2} s_n - \lambda_{2n} P\beta^{(0)}_n$, $s_n \in \partial\|\beta^{(0)}_n\|_1$, $\|s_n\| \le \sqrt{p}$. Since $h_n \to 0$ implies $\lambda_{2n}/\lambda^+_{\min}(n) \to 0$, applying $G_n^\dagger$ yields $\operatorname{dist}(\beta^{(0)}_n, [\theta]) = O(e_n)$, a.s.

Moreover, since $M_n = \sum_{k=1}^n \varphi_k w_{k+1} \in \operatorname{Range}(G_n)$, we have $d_n^\top M_n = (G_n^{1/2} d_n)^\top (G_n^{\dagger/2} M_n)$. Therefore, it follows

$$2d_n^\top M_n - d_n^\top G_n d_n = 2(G_n^{1/2} d_n)^\top (G_n^{\dagger/2} M_n) - \|G_n^{1/2} d_n\|^2 \le \|G_n^{\dagger/2} M_n\|^2 = O\big(\log\lambda_{\max}(n)\big),\ \text{a.s.}$$

We now explain how the two regularization terms select a representative from the equivalence class. The L_1-term first selects the minimum-L_1 representative set. From (46), we have

$$\lambda^{(1)}_{1n}\left(\|\beta^{(0)}_n\|_1 - \|\beta^\star_{1,P}\|_1\right) \le 2d_n^\top M_n - d_n^\top G_n d_n + \lambda_{2n}\left((\beta^\star_{1,P})^\top P\beta^\star_{1,P} - (\beta^{(0)}_n)^\top P\beta^{(0)}_n\right).$$

Since $\{\beta^{(0)}_n\}$ is bounded and $P \in \mathbb{S}^p_{++}$, the last term on the right-hand side is $O(\lambda_{2n})$. Therefore, $\left[\|\beta^{(0)}_n\|_1 - \|\beta^\star_{1,P}\|_1\right]_+ = O\left((\log\lambda_{\max}(n) + \lambda_{2n})/\lambda^{(1)}_{1n}\right)$, a.s. Since $\mathcal{S}_1 = \arg\min_{\beta\in[\theta]}\|\beta\|_1 = [\theta] \cap \{\beta : \|\beta\|_1 \le \|\beta^\star_{1,P}\|_1\}$, where $[\theta]$ is affine and $\{\beta : \|\beta\|_1 \le c_1\}$ is polyhedral, Hoffman's error bound [45] implies that

$$\operatorname{dist}(\beta^{(0)}_n, \mathcal{S}_1) = O\big(\operatorname{dist}(\beta^{(0)}_n, [\theta]) + [\|\beta^{(0)}_n\|_1 - \|\beta^\star_{1,P}\|_1]_+\big).$$

Hence, we get $\operatorname{dist}(\beta^{(0)}_n, \mathcal{S}_1) = O(h_n)$, a.s.

It remains to show that the L_2-term selects the minimum-P-norm element inside $\mathcal{S}_1$. Let $\widetilde{\beta}_n$ be the Euclidean projection of $\beta^{(0)}_n$ onto $[\theta]$. Since $\beta^\star_{1,P} \in \mathcal{S}_1$, it minimizes $\|\cdot\|_1$ over $[\theta]$. Hence $\|\beta^\star_{1,P}\|_1 \le \|\widetilde{\beta}_n\|_1$. Therefore, $\|\beta^\star_{1,P}\|_1 - \|\beta^{(0)}_n\|_1 \le \|\widetilde{\beta}_n\|_1 - \|\beta^{(0)}_n\|_1 \le \|\widetilde{\beta}_n - \beta^{(0)}_n\|_1 = O\big(\operatorname{dist}(\beta^{(0)}_n, [\theta])\big) = O(e_n)$, a.s. Returning to (46) and using the preceding estimates, we get $\left[(\beta^{(0)}_n)^\top P\beta^{(0)}_n - (\beta^\star_{1,P})^\top P\beta^\star_{1,P}\right]_+ = O\left(\frac{\log\lambda_{\max}(n)}{\lambda_{2n}} + \frac{\lambda^{(1)}_{1n}}{\lambda_{2n}} e_n\right)$, a.s.

Let z_n be the Euclidean projection of $\beta^{(0)}_n$ onto $\mathcal{S}_1$. Then $\|\beta^{(0)}_n - z_n\| = O(h_n)$, a.s. Since $\{\beta^{(0)}_n\}$ and $\{z_n\}$ are bounded, and P is fixed, $z_n^\top P z_n - (\beta^{(0)}_n)^\top P\beta^{(0)}_n = O(\|z_n - \beta^{(0)}_n\|) = O(h_n)$. Thus, $z_n^\top P z_n - (\beta^\star_{1,P})^\top P\beta^\star_{1,P} = O\big(h_n + \frac{\log\lambda_{\max}(n)}{\lambda_{2n}} + \frac{\lambda^{(1)}_{1n}}{\lambda_{2n}} e_n\big)$, a.s. Noting that $\mathcal{S}_1$ is convex and by (45), the first-order optimality condition gives

$2(\beta^\star_{1,P})^\top P(z_n-\beta^\star_{1,P}) \ge 0$. Hence $z_n^\top P z_n-(\beta^\star_{1,P})^\top P\beta^\star_{1,P} = (z_n-\beta^\star_{1,P})^\top P(z_n-\beta^\star_{1,P}) + 2(\beta^\star_{1,P})^\top P(z_n-\beta^\star_{1,P}) \ge \lambda_{\min}(P)\|z_n-\beta^\star_{1,P}\|^2$. Consequently,

$$\|z_n-\beta^\star_{1,P}\| = O\left(\sqrt{h_n+\frac{\log\lambda_{\max}(n)}{\lambda_{2n}}+\frac{\lambda^{(1)}_{1n}}{\lambda_{2n}}e_n}\right),\ \text{a.s.}$$

Finally, since $z_n, \beta^\star_{1,P} \in \mathcal{S}_1 \subset [\theta]$, we have $z_n-\beta^\star_{1,P} \in \mathcal{N}$. Thus, the error in the unidentifiable direction satisfies

$$\begin{aligned}\left\|\Pi_{\mathcal{N}}(\beta^{(0)}_n-\beta^\star_{1,P})\right\| &\le \|\beta^{(0)}_n-z_n\|+\|z_n-\beta^\star_{1,P}\| \\ &= O\left(\sqrt{h_n+\frac{\log\lambda_{\max}(n)}{\lambda_{2n}}+\frac{\lambda^{(1)}_{1n}}{\lambda_{2n}}e_n}\right),\ \text{a.s.}\end{aligned}$$

Combining this with the estimate on the identifiable component gives the desired bound for $\beta^{(0)}_n-\beta^\star_{1,P}$. ∎

Remark 13: For comparison, in the full-rank case $\mathcal{N}=\{0\}$, the same argument yields $[\theta]=\{\theta\}$, $\mathcal{S}_1=\{\theta\}$, $\beta^\star_{1,P}=\theta$, and $\Pi_{\mathcal{N}}=0$. Hence the result reduces to Corollary 1 with $\lambda^+_{\min}(n)=\lambda_{\min}(n)$.

We next derive the convergence and finite-step selection properties of Algorithm 1 under exact collinearity. After a possible permutation of coordinates, write $\beta^\star_{1,P} = ((\beta^\star_{1,P})^\top_{I^\star}, 0^\top_{|A^\star|})^\top$ with $I^\star=\{j:\beta^\star_{1,P}(j)\neq 0\}$ and $A^\star=\{j:\beta^\star_{1,P}(j)=0\}$. Define the limiting weighted representative set by $\mathcal{S}_{\rm w}=\arg\min_{\beta\in[\theta],\beta_{A^\star}=0}\sum_{j\in I^\star}\frac{|\beta(j)|}{|\beta^\star_{1,P}(j)|}$. Since $\beta^\star_{1,P}$ is feasible and the weights on $I^\star$ are positive, $\mathcal{S}_{\rm w}$ is nonempty, compact, and convex. Then we impose the following additional condition under Assumption 1.

(B1) There exist constants $0<c_i<C_i<\infty$, $i=1,2$, such that, for all sufficiently large n, $c_1 n\le\lambda_{\max}(n)\le C_1 n$, $c_2 n\le\lambda^+_{\min}(n)\le C_2 n$.

Theorem 5 (Convergence under exact collinearity): Suppose that Assumptions (A1)–(A2), Assumption 1, and condition (B1) hold. Let $\widehat{\beta}_n$ be generated by Algorithm 1 with $\lambda^{(1)}_{1n}=n^{1/2}$, $\lambda_{2n}=n^{1/4}$, $\lambda^{(2)}_{1n}=n^{1/2}$, and $\hat{\beta}^{(0)}_n(j)=|\beta^{(0)}_n(j)|+r_n$, where $r_1=1$, $r_n=n^{-1/8}\sqrt{\log n}$. Then $\mathrm{dist}(\widehat{\beta}_n,\mathcal{S}_{\rm w})=O(r_n)\to 0$, a.s. Moreover, there exists a set Ω_0 with $\mathbb{P}(\Omega_0)=1$ such that, for every $\omega\in\Omega_0$, there exists an integer $N_0(\omega)$ satisfying $A^\star\subseteq\widehat{A}^\star_n$ for all $n\ge N_0(\omega)$. If $\mathcal{S}_{\rm w}=\{\beta^\star_{\rm w}\}$, then $\widehat{\beta}_n\to\beta^\star_{\rm w}$, a.s.

Proof: By (B1) and Lemma 5, we have $\|\beta^{(0)}_n-\beta^\star_{1,P}\|=O(r_n)\to 0$, a.s. Hence, for $j\in I^\star$, $|\beta^{(0)}_n(j)|\ge\frac{1}{2}|\beta^\star_{1,P}(j)|$ eventually a.s. Thus, $\frac{1}{\hat{\beta}^{(0)}_n(j)}=\frac{1}{|\beta^{(0)}_n(j)|+r_n}=O(1)$, $j\in I^\star$. Write $b:=\beta^\star_{1,P}$ and $\omega_n(j):=1/\hat{\beta}^{(0)}_n(j)$, and define $Q_n(\beta):=L_n(\beta)+\lambda_{2n}\beta^\top P\beta$ and $\widetilde{\beta}_n:=\arg\min_{\beta_{A^\star}=0}\{Q_n(\beta)+\lambda^{(2)}_{1n}\sum_{j\in I^\star}\omega_n(j)|\beta(j)|\}$. Since the weights on $I^\star$ are bounded above and away from zero, comparison with b, as in (46), gives $\sum_{j\in I^\star}\omega_n(j)|\widetilde{\beta}_n(j)|\le\sum_{j\in I^\star}\omega_n(j)|b(j)|+O((\log n+\lambda_{2n})/\lambda^{(2)}_{1n})=O(1)$, and hence $\widetilde{\beta}_n=O(1)$, a.s. Let $K_n:=G_{n,I^\star I^\star}$. Its kernel is eventually fixed, and (B1) implies $\|K^\dagger_n\|=O(n^{-1})$. The restricted optimality condition gives $K_n(\widetilde{\beta}_{n,I^\star}-b_{I^\star})=M_{n,I^\star}-(\lambda^{(2)}_{1n}/2)v_n-\lambda_{2n}(P\widetilde{\beta}_n)_{I^\star}$, where $\|v_n\|=O(1)$. Since the zero extension of $\ker K_n$ is contained in $\mathcal{N}$, applying $K^\dagger_n$ and using $\|M_n\|=O(\sqrt{n\log n})$ yields $\|\Pi_{\mathcal{N}^\perp}(\widetilde{\beta}_n-b)\|=O(\sqrt{\log n/n})$, a.s. The case $I^\star=\varnothing$ is immediate.

Since $\sum_{k=1}^n\varphi_k y_{k+1}=G_n\theta+M_n$, and $G_n(\beta^\star_{1,P}-\theta)=0$, we have $\nabla Q_n(\beta)=2G_n(\beta-\beta^\star_{1,P})-2M_n+2\lambda_{2n}P\beta$. Moreover, since $\ker(G_n)=\mathcal{N}$, $G_n(\beta-\beta^\star_{1,P})=G_n\Pi_{\mathcal{N}^\perp}(\beta-\beta^\star_{1,P})$. Thus, the least-squares part of the gradient depends only on the identifiable component. For $j\in A^\star$, by Lemma 5, we have eventually a.s. $\frac{1}{\hat{\beta}^{(0)}_n(j)}=\frac{1}{|\beta^{(0)}_n(j)|+r_n}\ge\frac{c}{r_n}$, $j\in A^\star$. Using the bounds for $\widetilde{\beta}_n$, $\|\nabla Q_n(\widetilde{\beta}_n)\|\le 2\lambda_{\max}(n)\|\Pi_{\mathcal{N}^\perp}(\widetilde{\beta}_n-b)\|+2\|M_n\|+2\lambda_{2n}\|P\|\,\|\widetilde{\beta}_n\|=O(\sqrt{n\log n})$, a.s. Moreover, for $j\in A^\star$, $\lambda^{(2)}_{1n}\omega_n(j)\ge cn^{5/8}/\sqrt{\log n}$ eventually, a.s. Hence, since $n^{-1/8}\log n\to 0$, $|[\nabla Q_n(\widetilde{\beta}_n)]_j|<\lambda^{(2)}_{1n}\omega_n(j)$ eventually for all $j\in A^\star$, a.s.

Since $\widetilde{\beta}_{n,A^\star}=0$, the preceding strict inequalities imply that $\widetilde{\beta}_n$ satisfies the full subgradient optimality condition on $A^\star$. It already satisfies this condition on $I^\star$ by construction [46, Theorems 23.8 and 27.4]. By strong convexity, $\widehat{\beta}_n=\widetilde{\beta}_n$ eventually, a.s. Consequently, $A^\star\subseteq\widehat{A}^\star_n$ eventually, a.s.

Define $F_{\rm w}(\beta):=\sum_{j\in I^\star}|\beta(j)|/|b(j)|$, and fix $\gamma\in\mathcal{S}_{\rm w}$. For $j\in I^\star$, the first-stage estimate gives $\omega_n(j)-|b(j)|^{-1}=O(r_n)$. Since $\widehat{\beta}_n$ is bounded and $\widehat{\beta}_{n,A^\star}=0$ eventually, comparison with γ yields $[F_{\rm w}(\widehat{\beta}_n)-F_{\rm w}(\gamma)]_+=O(r_n+(\log n+\lambda_{2n})/\lambda^{(2)}_{1n})=O(r_n)$. Hence, by Hoffman's error bound [45], $\mathrm{dist}(\widehat{\beta}_n,\mathcal{S}_{\rm w})\le C\{\|\Pi_{\mathcal{N}^\perp}(\widehat{\beta}_n-b)\|+\|\widehat{\beta}_{n,A^\star}\|+[F_{\rm w}(\widehat{\beta}_n)-F_{\rm w}(\gamma)]_+\}=O(r_n)$, a.s. If $\mathcal{S}_{\rm w}=\{\beta^\star_{\rm w}\}$, then $\|\widehat{\beta}_n-\beta^\star_{\rm w}\|=O(r_n)\to 0$, a.s.

∎

Remark 14: Under the assumptions and parameter choices of Theorem 5, the first-stage estimator $\beta^{(0)}_n$ converges, a.s., to the unique representative $\beta^\star_{1,P}=\arg\min_{\beta\in\mathcal{S}_1}\beta^\top P\beta$ with $\mathcal{S}_1=\arg\min_{\beta\in[\theta]}\|\beta\|_1$. This first-stage selection mechanism is illustrated below: asymptotically, the L_1 term selects the minimum-L_1 representative set $\mathcal{S}_1$, and the quadratic P-term selects its unique minimum-P-norm element.

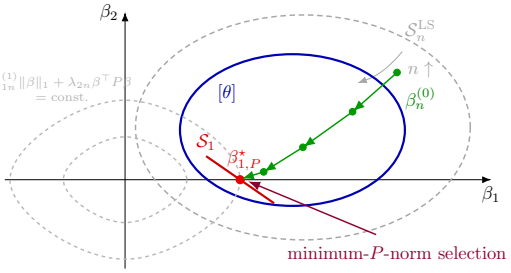

Fig. 1: Geometric illustration of the first-stage regularized selection mechanism under exact collinearity.

The first-stage limit determines the adaptive weights. The second-stage estimator satisfies $\mathrm{dist}(\widehat{\beta}_n,\mathcal{S}_{\rm w})\to 0$, a.s., where $\mathcal{S}_{\rm w}=\arg\min_{\substack{\beta\in[\theta]\\ \beta_{A^\star}=0}}\sum_{j\in I^\star}\frac{|\beta(j)|}{|\beta^\star_{1,P}(j)|}$. In general, $\mathcal{S}_{\rm w}$ need not equal $\{\beta^\star_{1,P}\}$. Point convergence follows when $\mathcal{S}_{\rm w}$ is a singleton, while $A^\star\subseteq\widehat{A}^\star_n$ holds eventually, a.s.

In addition, the LS solution set admits a set-valued interpretation. For $G\in\mathbb{S}^p_+$ and $M\in\mathrm{range}(G)$, define $\mathcal{F}(G,M)=\{\beta\in\mathbb{R}^p: G\beta=G\theta+M\}$. Then $\mathcal{S}^{\rm LS}_n=\mathcal{F}(G_n,M_n)=G^\dagger_n(G_n\theta+M_n)+\ker(G_n)$ is nonempty, closed, and convex. On any parameter domain where $\mathcal{F}$ is Hausdorff-continuous,

its minimum-P-norm selection is continuous, viewing $\mathbb{R}^p$ as a Hilbert space with norm $\|x\|_P = (x^\top P x)^{1/2}$; see [47, Section 1.7].

Remark 15 (From Algorithm 1 to L_0-sparsity): The proposed estimator does not directly solve the combinatorial L_0 problem, and $\mathcal{S}_{\mathrm{w}}$ need not consist of L_0-sparsest representatives. Under the assumptions and parameter choices of Theorem 5, the following condition guarantees convergence to an L_0-sparsest representative. Suppose that there exists $\beta^0 \in \mathcal{S}_0$ such that, with $T = \mathrm{supp}(\beta^0)$, $\|h_T\|_1 < \|h_{T^c}\|_1$ for all $h \in \mathcal{N} \setminus \{0\}$. Then $\|\beta^0 + h\|_1 \geq \|\beta^0\|_1 - \|h_T\|_1 + \|h_{T^c}\|_1 > \|\beta^0\|_1$, so that $\mathcal{S}_1 = \{\beta^0\}$ and $\beta^\star_{1,P} = \beta^0$. Moreover, this condition implies $\mathcal{N} \cap \{h : h_{T^c} = 0\} = \{0\}$. Since $I^\star = T$ and $A^\star = T^c$, we have $\mathcal{S}_{\mathrm{w}} = [\theta] \cap \{\beta : \beta_{T^c} = 0\} = \{\beta^0\}$. Consequently, $\widehat{\beta}_n \to \beta^0 \in \mathcal{S}_0$ a.s., and $\mathrm{supp}(\widehat{\beta}_n) = T$ eventually a.s. Related null-space conditions for sparse recovery are discussed in [48, Sections 3–4].

V. Simulation Study

This section presents three numerical simulations and a real-world example to evaluate the performance of Algorithm 1. Since the objective function (12) is strictly convex, we employ the CVX toolbox for MATLAB to compute the minimizer in (13). All numerical tests are conducted in MATLAB R2025a on a Lenovo desktop (2.60 GHz, 32 GB RAM).

Example 1. This example compares Algorithm 1 with LS and the weighted L_1 method in [20] under correlated regressors. Consider the system (1) with $\{\varphi_k\}$ being i.i.d. Gaussian vectors with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$, and $\{w_k\}$ being i.i.d. $N(0,1)$ and independent of $\{\varphi_k\}$. To introduce collinearity, we take $\sigma_{ij} = \kappa^{|i-j|}$, $\kappa = 0.7$.

We consider $(p,q) = (15,10)$ and $(p,q) = (40,15)$, where q is the number of nonzero components of θ. The regularization parameters are chosen as follows: for the method in [20], $\lambda_n = n^{0.75}$; for Algorithm 1, $\lambda_{1n}^{(1)} = \lambda_{1n}^{(2)} = n^{0.75}$, $\lambda_{2n} = n^{0.15}$. The performance is measured by the MSE, the probability of correctly selecting zero coefficients (PCS), and the probability of incorrectly selecting nonzero coefficients (PIS). Each experiment is repeated 50 times.

Table I shows that both sparse methods select zero coefficients effectively, while Algorithm 1 usually achieves smaller MSE, especially for larger p. This suggests that the quadratic term improves stability under correlated regressors.

Example 2. This example illustrates the group-selection property of Algorithm 1. Consider the multi-group FIR model $y_{k+1} = \sum_{g=1}^{G} \theta_g^\top \varphi_k^{(g)} + w_{k+1}$, where $G = 10$ and $\sum_{g=1}^{G} p_g = 100$. The coefficients in active groups are nonzero and have the same sign, while inactive groups have zero coefficients. For each group, $\varphi_k^{(g)} = \sqrt{\rho}\, b_k^{(g)} \mathbf{1}_{p_g} + \sqrt{1-\rho}\, \epsilon_k^{(g)}$, $\rho = 0.99$, where $b_k^{(g)} \sim N(0,1)$ and $\epsilon_k^{(g)}$ has i.i.d. $N(0,1)$ entries. Regressors from different groups are independent, and $w_k \sim N(0, 1.5^2)$. We set $N = 150$ and repeat the experiment 100 times. Algorithm 1 is compared with group LASSO with the correct group partition, group LASSO with random groups, weighted L_1 [20], and LASSO.

TABLE I: Estimation error and selection performance under $\sigma_{ij} = \kappa^{|i-j|}$, $\kappa = 0.7$.

p	n	q	Method	MSE	PCS	PIS
15	100	10	LS	0.248 (0.428)	0	0
			Algorithm in [20]	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 [20]	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 [20]	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 [20]	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 [20]	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 [20]	0.271 (0.357)	0.999	0.001
			Algorithm 1	0.256 (0.169)	0.992	0

Fig. 2 reports boxplots of selection accuracy and the empirical cumulative distribution functions (ECDFs) of the L_2 estimation errors. Algorithm 1 achieves more stable selection and smaller errors, since it tends to retain highly correlated variables within the same active group.

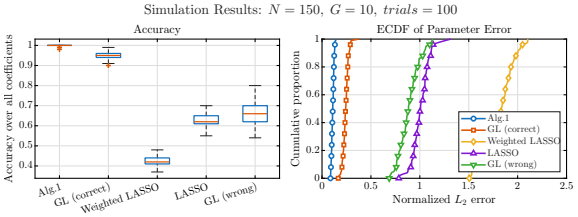

Fig. 2: Comparative results in Example 2: boxplots of selection accuracy (top) and ECDFs of estimation errors (bottom).

Example 3. This example illustrates the convergence, finite-step selection, stability, and robustness of Algorithm 1 under exact collinearity. Consider system (1) with $p = 240$. The parameter vector is divided into $d = 60$ groups, each containing four components. For each $\ell = 1, \ldots, 60$, the regressors in the same group are identical: $\varphi_{k,4\ell-3} = \varphi_{k,4\ell-2} = \varphi_{k,4\ell-1} = \varphi_{k,4\ell} = z_{k,\ell}$. The vector $z_k = (z_{k,1}, \ldots, z_{k,60})^\top$ is Gaussian with Toeplitz covariance matrix $\mathrm{Cov}(z_{k,i}, z_{k,j}) = \rho^{|i-j|}$, $\rho = 0.35$. Thus only the group sums $q_\ell = \beta_{4\ell-3} + \beta_{4\ell-2} + \beta_{4\ell-1} + \beta_{4\ell}$, $\ell = 1, \ldots, 60$, are identifiable. With the group-sum matrix $A \in \mathbb{R}^{60 \times 240}$, the identifiable equivalence class is $[\theta] = \{\beta \in \mathbb{R}^{240} : A\beta = s\}$.

The group-sum vector s is sparse. Its active groups are $\{1, \ldots, 12\} \cup \{18, \ldots, 30\} \cup \{38, \ldots, 49\}$, and all other group sums are zero. The observation noise is i.i.d. Gaussian with standard deviation 0.18. Since the original parameter is not uniquely identifiable, we use the selected representative $\beta^*_{1,P} = \arg\min_{\beta \in \mathcal{S}_1} \beta^\top P \beta$, $\mathcal{S}_1 = \arg\min_{\beta \in [\theta]} \|\beta\|_1$, as the first-stage target. The second-stage estimator approaches the weighted-L_1 solution set S_w. Here $P \in \mathbb{S}^{240}_{++}$ specifies the geometry used to select a unique representative

from the minimum-L_1-norm set. Four fixed choices of P are tested: (i) a diagonal matrix with entries sampled from $[0.6, 1.4]$; (ii) a block diagonal matrix whose 4×4 blocks are $a_\ell(U_\ell^\top U_\ell/4 + 0.4I_4)$, with $a_\ell \in [0.8, 1.2]$; (iii) a Toeplitz matrix $(0.72^{|i-j|})_{ij} + 0.3I_p$; and (iv) a full matrix $W^\top W/p + 0.6I_p$, where the entries of U_ℓ and W are standard Gaussian. These choices cover diagonal, group-structured, globally correlated, and fully dense geometries. The regularization parameters are chosen as $\lambda_{1n}^{(1)} = 0.10\, n^{0.5}$, $\lambda_{2n} = 0.25\, n^{0.25}$, $\lambda_{1n}^{(2)} = 0.10\, n^{0.5}$ and the maximum sample size is $N = 20000$. The estimator is computed by FISTA [49]. The target $\beta_{1,P}^*$ is computed by solving the corresponding quadratic minimization problem over $\mathcal{S}_1$, where the projection onto the product of simplexes is implemented by the sorting-based method in [50].

Fig. 3 illustrates the conclusions of Theorem 5. The first panel shows the convergence of $\beta_n^{(0)}$ to $\beta_{1,P}^*$ for all four choices of P. The second panel reports the finite recovery sample size, defined as the first n such that $\#\left\{j : \beta_{1,P}^*(j) = 0,\ |\widehat{\beta}_n(j)| > 3.0 \times 10^{-3} n^{-0.12}\right\} = 0$. Thus, the first stage estimate converges to $\beta_{1,P}^*$, while the second stage estimate recovers the zero coordinates of $\beta_{1,P}^*$ in finite samples.

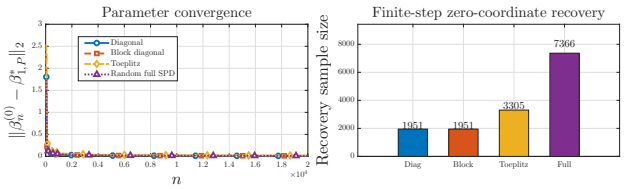

Fig. 3: Theory validation under exact collinearity.

We next examine the stability mechanism discussed in Remark 14. Using the same exact-collinearity model, let $\mathbf{X} = (\varphi_1, \ldots, \varphi_N)^\top \in \mathbb{R}^{N \times p}$, $\mathbf{y} = (y_2, \ldots, y_{N+1})^\top \in \mathbb{R}^N$. We take $N = 6000$ and add small independent perturbations to both $\mathbf{X}$ and $\mathbf{y}$: $\mathbf{X}^{(r)} = \mathbf{X} + 2.0 \times 10^{-3} E^{(r)}$, $\mathbf{y}^{(r)} = \mathbf{y} + 2.0 \times 10^{-3} \operatorname{std}(\mathbf{y}) e^{(r)}$, where the entries of $E^{(r)} \in \mathbb{R}^{N \times p}$ and $e^{(r)} \in \mathbb{R}^N$ are i.i.d. standard normal random variables. The experiment is repeated 50 times.

We compare LS, LASSO, adaptive LASSO, and the proposed method with block diagonal, Toeplitz, and random full positive definite matrices P. Fig. 4 further confirms the stabilizing effect of the quadratic P-term. In the left panel, the vertical axis is in logarithmic scale. The LS, LASSO, and adaptive LASSO estimators are sensitive to small input/output perturbations, with average coordinate-wise variances at the order of 10^1. In contrast, the proposed estimators remain stable, with variances below 10^{-5}. The right panel compares sparsity and support-size stability. LS is stable in support size but is not sparse. LASSO produces sparse estimates, but its support size varies noticeably under perturbations. Adaptive LASSO is overly sparse in this example. The proposed estimators achieve both sparsity and stability, showing that the L_1-term preserves sparse selection while the quadratic P-term stabilizes the selected representative.

We further evaluate the robustness of the proposed method under different SNRs, sparsity levels, and degrees of collinearity. Since LS gives much larger prediction errors in these

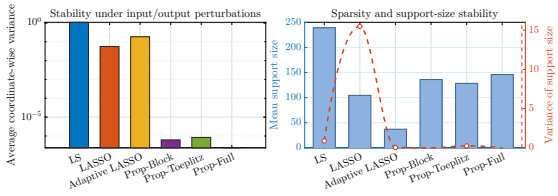

Fig. 4: Stability and sparsity under small perturbations.

strongly collinear settings, it is omitted from Fig. 5. We compare LASSO, adaptive LASSO, and the proposed estimator with block diagonal and random full positive definite matrices P. The training and test sample sizes are $N_{\mathrm{tr}} = 360$ and $N_{\mathrm{te}} = 800$, respectively, and each experiment is repeated 25 times. The regressors are generated from a near-collinear group model. Prediction performance is measured by the normalized test error $\mathrm{NMSE} = \dfrac{N_{\mathrm{te}}^{-1} \sum_{i=1}^{N_{\mathrm{te}}} \left(x_i^\top \widehat{\beta} - x_i^\top \beta\right)^2}{N_{\mathrm{te}}^{-1} \sum_{i=1}^{N_{\mathrm{te}}} \left(x_i^\top \beta\right)^2}$. For LASSO, we set $\lambda = 1.10\, \sigma \sqrt{N_{\mathrm{tr}} \log p}$. For adaptive LASSO, the initial LASSO estimate is computed with 0.55λ, and the adaptive weights are $\omega_j = \frac{1}{|\beta_n^{(0)}(j)| + N_{\mathrm{tr}}^{-1/4}}$. For the proposed method, we use $\lambda_1^{(1)} = 0.55\lambda$, $\lambda_1^{(2)} = 0.48\lambda$, $\lambda_2 = 0.10\, N_{\mathrm{tr}}$.

Fig. 5 reports the NMSE. The variants of Algorithm 1, with different choices of P, show comparatively stable performance across the tested settings. As the SNR increases, all methods improve and the performance gap becomes smaller. As the number of active groups increases, the problem becomes more difficult, while Algorithm 1 remains competitive. For mild within-group collinearity, LASSO performs well; however, its error increases more evidently when collinearity becomes stronger. In contrast, the variants of Algorithm 1 are less sensitive to the collinearity level. These results show the robustness of Algorithm 1 in ill-conditioned settings, as well as its limitations in easier regimes.

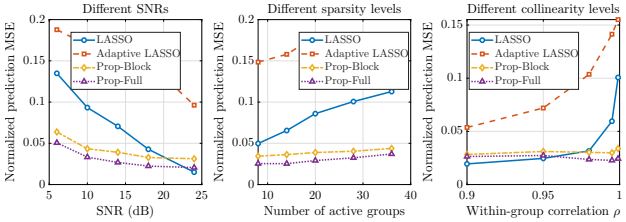

Fig. 5: Robustness comparison in normalized prediction MSE.

Example 4. A high-fidelity robot model is important for precise positioning and tracking control. Following [51], we use the vibrating flexible robot arm dataset, which has also been studied in [24], [52]. The dataset contains 40,960 samples collected at 500 Hz, where the input is the driving torque and the output is the acceleration measured at the arm tip.

We compare Algorithm 1 with the methods in [20], [24], [52]. Since the true system is unknown, we use hold-out validation: samples 1:7000 are used for estimation, and samples 10,000:40,960 are used for validation. All identified models have order 2500. The performance is measured by the FIT index $\mathrm{FIT} = 100\left(1 - \dfrac{\|y - \hat{y}\|}{\|y - \mathrm{mean}(y)\|}\right)$, where y is the measured output and $\hat{y}$ is the simulated output. A larger FIT

indicates better simulation accuracy.

Table II reports the validation FIT for five variants of Alg. 1, corresponding to five choices of the quadratic penalty matrix P_n: (I) $P_n = I_{57}$; (II) $P_n = \mathrm{diag}\{1.8I_{45}, 0.3I_{12}\}$; (III) $P_n = D^\top D + \alpha I_{57}$; (IV) $P_n = (1-\alpha)L/\lambda_{\max}(L) + \alpha I_{57}$; and (V) $P_n = (1-\rho)K_n + \rho\mu I_{57}$. Here D is the first-order difference matrix, L is the block Laplacian matrix, $K_n = n^{-1}\sum_{k=1}^{n}\varphi_k\varphi_k^\top$, $\mu = p^{-1}\,\mathrm{tr}(K_n)$, and $\alpha, \rho > 0$ are chosen so that P_n is positive definite.

We also compare with a post-LS refitting version of [20], where LS is performed on the support selected by [20]. Moreover, Fig. 6 shows the top 100 coefficients. Alg. 1 (V) concentrates most of them in the lag range 0–600, while the other methods either miss some coefficients in this range or spread weight to longer lags. This concentration is consistent with the short-memory behavior of the system.

TABLE II: 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%
[20] + LS refitting	85.042%
[20]	83.988%
[52]	80.100%
[24]	79.900%

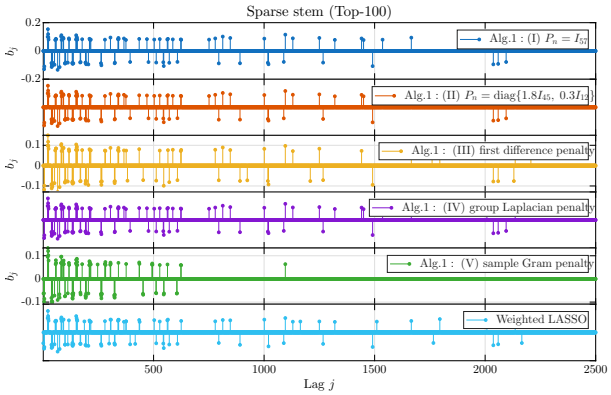

Fig. 6: Sparse stem plots of the top-100 coefficients for six algorithms.

VI. Conclusion

This paper presents a regularized sparse identification framework for non-i.i.d. and non-stationary observations. The weighted L_1 term promotes sparse selection, while the quadratic term improves stability under collinearity and induces the group effect. We establish almost sure parameter convergence, finite-step set convergence, and asymptotic normality. We also study exact collinearity, where the original parameter vector is not uniquely identifiable. The first stage converges to the minimum-P-norm element of the minimum-L_1 set, while the second stage converges to the corresponding weighted L_1 solution set with finite-step zero recovery. Future work includes high-dimensional settings with $p = p(n)$ and recursive algorithms for real-time identification and control.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

[7] D. Kepplinger, "Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression," *Comput. Statist. Data Anal.*, vol. 183, p. 107730, 2023.

[8] A. D. Gregorio, D. Frisardi, S. Iacus, and F. Iafrate, "Adaptive elastic-net estimation for sparse diffusion processes," *Statist. Inference Stoch. Process.*, vol. 28, p. 22, 2025.

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

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

[11] 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.

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

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

[14] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Coherence-based performance guarantees for estimating a sparse vector under random noise," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5030–5043, 2010.

[15] J. Parsa, C. R. Rojas, and H. Hjalmarsson, "Balancing application relevant and sparsity revealing excitation in input design," *IEEE Trans. Autom. Control*, vol. 70, no. 3, pp. 1890–1897, Mar. 2025.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

[28] 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.

[29] 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.

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

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

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

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

[34] 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.

[35] 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.

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

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

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] A. J. Hoffman, "On approximate solutions of systems of linear inequalities," *Journal of Research of the National Bureau of Standards*, vol. 49, no. 4, pp. 263–265, 1952.

[46] R. T. Rockafellar, *Convex Analysis*, ser. Princeton Mathematical Series. Princeton, NJ: Princeton University Press, 1970, vol. 28.

[47] J.-P. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*. Berlin: Springer, 1984.

[48] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American Mathematical Society*, vol. 22, no. 1, pp. 211–231, 2009.

[49] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[50] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 272–279.

[51] 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.

[52] 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 6: 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

$$J_n\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n),\alpha_{1n},\alpha_{2n},\eta_n\big) \le J_n(\theta,\alpha_{1n},\alpha_{2n},\eta_n). \quad (47)$$

Using (1) and (14), and noting that $\theta(j) = 0$ for $j = q+1,\dots,p$, we obtain

$$\begin{aligned} &J_n(\theta,\alpha_{1n},\alpha_{2n},\eta_n)\\ &=\sum_{k=1}^{n} w_{k+1}^2 + \alpha_{1n}\sum_{j=1}^{q}\eta_n(j)\,|\theta(j)| + \alpha_{2n}\,\theta^\top P_n\theta, \end{aligned} \quad (48)$$

$$\begin{aligned} &J_n\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n),\alpha_{1n},\alpha_{2n},\eta_n\big)\\ &=\sum_{k=1}^{n} w_{k+1}^2 + \sum_{k=1}^{n}\big((\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta)^\top\varphi_k\big)^2\\ &\quad - 2\,(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta)^\top\sum_{k=1}^{n}\varphi_k w_{k+1}\\ &\quad + \alpha_{1n}\sum_{j=1}^{p}\eta_n(j)\,\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)\big|\\ &\quad + \alpha_{2n}\,\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)^\top P_n\,\beta_n(\alpha_{1n},\alpha_{2n},\eta_n), \end{aligned} \quad (49)$$

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 (49)–(48), we have

$$\begin{aligned} &J_n\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n),\alpha_{1n},\alpha_{2n},\eta_n\big) - J_n(\theta,\alpha_{1n},\alpha_{2n},\eta_n)\\ &\ge\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big)^\top\sum_{k=1}^{n}\varphi_k\varphi_k^\top\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big)\\ &\quad + \alpha_{2n}\left[\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)^\top P_n\beta_n(\alpha_{1n},\alpha_{2n},\eta_n) - \theta^\top P_n\theta\right]\\ &\quad - 2\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big)^\top\sum_{k=1}^{n}\varphi_k w_{k+1}\\ &\quad + \alpha_{1n}\sum_{j=1}^{q}\eta_n(j)\left(\big|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)(j)\big| - |\theta(j)|\right)\\ &\triangleq M_n^{(1)} + M_n^{(2)} - 2M_n^{(3)} + M_n^{(4)}. \end{aligned} \quad (50)$$

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

$$\delta_n = \left(\sum_{k=1}^{n}\varphi_k\varphi_k^\top + \alpha_{2n}P_n\right)^{\frac{1}{2}}\big(\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\big). \quad (51)$$

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

$$\begin{aligned} &M_n^{(1)} + M_n^{(2)}\\ &\ge\delta_n^\top\delta_n - 2\alpha_{2n}\lambda_{\max}(P_n)\,\|\theta\|\,\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|. \end{aligned} \quad (52)$$

Second, by Lemma 6, we have $\|(\sum_{k=1}^{n}\varphi_k\varphi_k^\top + \alpha_{2n}P_n)^{-\frac{1}{2}}(\sum_{k=1}^{n}\varphi_k\varphi_k^\top)^{\frac{1}{2}}\| \le 1$. Hence,

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

Thus, for some $c_1 > 0$ and all sufficiently large n,

$$-M_n^{(3)} \ge -\frac{c_1}{2}\sqrt{\log\lambda_{\max}(n)}\,\|\delta_n\|. \tag{53}$$

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

$$|M_n^{(4)}| \le \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}\,\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|. \tag{54}$$

Substituting (47), (52), (53), and (54) into (50) yields

$$\begin{aligned}0 \ge\ &\|\delta_n\|^2 - \alpha_{2n}\lambda_{\max}(P_n)\|\theta\|\,\|\beta_n-\theta\| \\ &- c_1\sqrt{\log\lambda_{\max}(n)}\,\|\delta_n\| \\ &- \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}\,\|\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)-\theta\|.\end{aligned} \tag{55}$$

By solving the quadratic inequality (55) with respect to $\|\delta_n\|$ and using $(a+b)^2 \le 2a^2+2b^2$, we obtain $\|\delta_n\|^2 \le c_1^2\log\lambda_{\max}(n) + 2\alpha_{2n}\lambda_{\max}(P_n)\|\theta\|\,\|\beta_n-\theta\| + 2\alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}\,\|\beta_n-\theta\|$. where, for brevity, β_n denotes $\beta_n(\alpha_{1n},\alpha_{2n},\eta_n)$. Moreover, recalling (51) and using $\lambda_{\min}(\sum_{k=1}^{n}\varphi_k\varphi_k^\top + \alpha_{2n}P_n) \ge \lambda_{\min}(n) + \alpha_{2n}\lambda_{\min}(P_n)$, we have $\|\delta_n\|^2 \ge \big(\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)\big)\,\|\beta_n-\theta\|^2$. Combining these yields

$$\begin{aligned}&\left(\|\beta_n-\theta\| - \frac{\alpha_{2n}\lambda_{\max}(P_n)\|\theta\| + \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\right)^2 \\ &\le \left(\frac{\alpha_{2n}\lambda_{\max}(P_n)\|\theta\| + \alpha_{1n}\sqrt{\sum_{j=1}^{q}\eta_n(j)^2}}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}\right. \\ &\quad\left.+\sqrt{\frac{c_1^2\log\lambda_{\max}(n)}{\lambda_{\min}(n)+\alpha_{2n}\lambda_{\min}(P_n)}}\right)^2.\end{aligned}$$

Taking square roots and combining the two terms on the right-hand side yields the asserted estimate. ■

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