Localization and Estimation of Unknown Forced Inputs: A Group LASSO Approach

Rajasekhar Anguluri, *Member, IEEE,* Lalitha Sankar, *Senior Member, IEEE,* and Oliver Kosut, *Member, IEEE*

***Abstract*—This paper studies the problem of locating the sparse set of sources of forcing inputs driving linear systems from noisy measurements when the initial state is unknown. This problem is particularly relevant to detecting forced oscillations in electric power networks. We express measurements as an additive model comprising the initial state and inputs grouped over time, both expanded in terms of the basis functions (i.e., impulse response coefficients). Using this model, with probabilistic guarantees, we recover the locations and simultaneously estimate the initial state and forcing inputs using a variant of the group LASSO (linear absolute shrinkage and selection operator) method. Specifically, we provide a tight upper bound on: (i) the probability that the group LASSO estimator wrongly identifies the source locations and (ii) the ℓ_2-norm of the estimation error. Our bounds depend on the length of the measurement horizon, the noise statistics, the number of inputs and sensors, and the minimum singular value of the impulse response matrices. Our theoretical analysis is one of the first to provide a complete treatment for the group LASSO estimator for the left invertible linear systems. Finally, we validate the performance of our estimator on synthetic models and the IEEE 68-bus, 16-machine system.**

***Index Terms*—Forced oscillations, unknown input, group LASSO, invariant zeros, source localization, sparse estimation.**

I. Introduction

Low-frequency oscillations in the electric transmission grid are indicative of the type of disturbance afflicting the system. *Natural oscillations*, with frequencies in between 0.1–2 Hz, are triggered by random load fluctuations and sudden network switching. In contrast, *forced oscillations* (FOs), with frequencies in between 0.1–14 Hz, result from external inputs injected by malfunctioned devices, such as power system stabilizers (PSS), generator controllers and exciters, and cyclic loads etc. [1], [2]. FOs remain undamped for longer periods of time, and if not mitigated, they pose a greater risk to the power systems operation, potentially causing blackouts.

A popular and an inexpensive method adopted in practice to mitigate FOs in power systems is to disconnect the sources triggering the oscillations [2]–[4]. This amounts to accurately locating the FO sources. As installing sensors at each potential source is expensive, recent research suggests using phasor measurement unit (PMU) measurements based source localization algorithms. These algorithms range from physics-based energy approaches to completely data-driven approaches [2]; the latter, albeit their impressive performance on test cases, lack theoretical guarantees. This deficiency makes it harder to quantify the performance and limitations of measurement-based methods on what is and is not possible.

We address the lack of guarantees of existing data-driven approaches by posing the localization problem as a regularized optimization problem—referred to as the group LASSO estimator. The regularization term imposes sparsity constraints on the number of source locations, which is often the case in many practical systems, including power systems [3], [5]. The input to our optimization problem are the noisy measurements and dynamical system matrices. It returns the source locations and estimates of unknown initial state and inputs (oscillatory or not) injected by these sources. Formally, we consider

$$\underbrace{\begin{bmatrix}\widehat{\mathbf{x}}_0\\ \widehat{\mathbf{u}}\end{bmatrix}}_{\widehat{\boldsymbol{\beta}}} \in \operatorname*{arg\,min}_{\mathbf{x}_0,\{\mathbf{u}_j\}_{j=1}^m} \left\Vert \mathbf{y}-\mathbf{O}\mathbf{x}_0-\sum_{j=1}^{m}\mathbf{J}_j\mathbf{u}_j\right\Vert_2^2+\lambda\sum_{j=1}^{m}\Vert\mathbf{u}_j\Vert_2\,, \quad (1)$$

where $\mathbf{u}_j=[u_j[0],\ldots,u_j[N]]^\mathsf{T}$ is a vector of inputs injected by the j^{th} source, $j\in\{1,\ldots,m\}$, over a discrete time horizon $\{0,\ldots,N\}$; $\mathbf{y}$ is the noisy batch measurements collected by p sesnors over $\{0,\ldots,N\}$; $\mathbf{O}$ and $\mathbf{J}_j$ are the observability and forced impulse response matrices, respectively (see Section II for the actual expressions); and $\lambda\geq 0$ is the tuning parameter. Let $\boldsymbol{\beta}^*=(\mathbf{x}_0^*,\mathbf{u}_1^*,\ldots,\mathbf{u}_m^*)$ be the unknown ground truth and $S\triangleq\{j:\mathbf{u}_j^*\neq 0\}\subset\{1,\ldots,m\}$ be the set of active sources. By sparse number of sources, we mean that $|S|=m^*\ll m$. Let $\widehat{S}\triangleq\{j:\widehat{\mathbf{u}}_j\neq 0\}$, where $\widehat{\mathbf{u}}$ is in (1). We show that $\widehat{S}=S^*$ and $\Vert\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\Vert_2\leq\epsilon$, for $\epsilon>0$, hold with high probability.

In the context of regression models, including linear, logistic, and functional models, a vast body of literature exists on quantifying the theoretical performance of the group LASSO estimator and its variants; for a sample, see [6]–[8]. However, these studies assume either $\mathbf{J}_j$ and $\mathbf{O}$ to be random or satisfy rather restrictive assumptions, both of these might not hold for $\mathbf{J}_j$ and $\mathbf{O}$ obtained from linear dynamical systems. Further, $\mathbf{J}_j$ associated with the non-zero input $\mathbf{u}_j^*$ could be rank deficient, especially if the underlying linear dynamical system is only d-*delay* left invertible[1] [9]. Consequently, the optimization in (1) is not strictly convex in the optimization variables, even when the true sources set S is known. Thus, there may exist multiple optimal solutions $\widehat{\boldsymbol{\beta}}$, and it is not clear if $\widehat{S}$ is common for all these solutions. In this paper, we address all these issues by imposing physically meaningful assumptions on $\mathbf{O}$ and $\mathbf{J}_i$.

This material is based upon work supported by the National Science Foundation under Grant No. OAC-1934766, and the U.S.-Israel Energy Center managed by the Israel-U.S. Binational Industrial Research and Development (BIRD) Foundation. The authors are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: {rangulur,lalithasankar,okosut}@asu.edu).

[1] A dynamical system is d-delay left invertible if $u_j[k]$, for $j\in\{1,\ldots,m\}$, can be uniquely recovered from noise-less measurements $\{y[k],\ldots,y[k+d]\}$.

Going beyond the motivating example of forced oscillations in electric power systems, the problem setup in (1) is general and the formal results in this paper can be used to localize and reconstruct sparse inputs for a variety of practical engineering systems modeled as linear dynamical systems.

Paper Contributions: The problem we introduce in (1) is distinct from state of the art regularized based optimization methods in seeking to localize inputs and estimate initial state using sufficiently delayed measurements over a block of time. For the estimator in (1), our main contributions as follows:

1) We derive sufficient conditions under which the following hold with high probability: (i) the estimation error in the ℓ_2-sense is bounded, and (ii) the localized sources match the true sources. A key contribution is that despite the rank deficiency of model matrices, we guarantee that the group LASSO can localize the sources correctly. For rank deficient matrices $\mathbf{J}_i$, we provide estimation guarantees for the delayed inputs (see Section III). Our result hinges on introducing and thresholding a *mutual incoherence condition* (MIC) on the augmented $\mathbf{O}$ and $\mathbf{J}_i$ matrices.
2) The time-domain MIC condition we introduce requires computing correlations among the matrices $\mathbf{O}$ and $\mathbf{J}_i$. This computation can be challenging especially for estimation horizon N. To tackle this hurdle, we upper bound the time-domain MIC with a frequency-domain MIC. Interestingly, the latter is a sufficient condition if we were to consider a LASSO estimator in the frequency-domain. We also establish a fundamental relationship between the performance of the proposed estimator and the absence of invariant zeros for the sub-system excited by non-zero inputs, and thresholding the frequency domain MIC.
3) We validate the group LASSO estimator's performance on synthetic data and the IEEE 68-bus, 16-machine system. We implement our estimator using the Alternating Direction Method Multipliers (ADMM) method [10].

Related Literature: In power systems, energy methods based on frequency domain data and statistical signal processing methods (e.g. AR and ARMA models) are commonly used to localize unknown forced oscillatory inputs. In [11], a Bayesian approach was used to localize sources based on the generators frequency response functions. In [12], the pseudo-inverse of system transfer functions were used to localize the sources. In [13], the authors leveraged magnitude and phase responses of transfer functions between different buses to localize sources. Finally, machine learning and PCA methods for localization were explored in [14] and [3]. All these methods primarily focus on oscillatory inputs, which might not apply to non-oscillatory inputs, including malicious attacks. We address this limitation by casting the source localization problem as an unknown input recovery in linear dynamical systems.

The problem of source identification in the context input and sensor and attacks was studied in [15]–[18]. But they fail to address input estimation and are applicable only for noise-free systems with more sensors than inputs. Moreover, these methods rely on banks of input observers, which might not be a practical solution for large-scale critical infrastructures. In [19], [20], for known inputs, using randomly sampled measurements, the authors obtained sample complexity results for reconstructing the initial state with sparsity constraints. Instead, the authors in [21] and [22] considered sparse input and non-sparse state reconstruction using off-line and sequential measurements. However, these works do not address location recovery guarantees for the sparse set of unknown sources in the presence of an unknown non-sparse initial state. In contrast to these works, we consider a unified framework, based on a LASSO method, to jointly locate the sources, and estimate the sparse inputs along with the unknown initial state. As highlighted in several other non-sparsity based input identification methods [15], [23], [24], our results also highlight the role of invariant zeros for sparse input recovery.

Mathematical Notation: We denote the vectors and matrices are by boldface lower case and upper case letters. Denote the $d \times d$ identity matrix by $\mathbf{I}_d$. When the context is clear we drop the subscript notation. Denote the pseudoinverse of $\mathbf{X}$ by $\mathbf{X}^{\dagger}$. The rangespace of $\mathbf{X}$ is defined by $\mathcal{R}(\mathbf{X}) = \{\mathbf{X}\mathbf{z} : \mathbf{z} \in \mathbb{R}^m\}$. Given $S \subset \{1, \ldots, m\}$ and $\mathbf{x} \in \mathbb{R}^m$, we write $\mathbf{x}_S$ for the sub-vector of $\mathbf{x}$ formed from the entries of $\mathbf{x}$ indexed by S. Similarly, we write $\mathbf{M}_S$ for the submatrix of $\mathbf{M}$ formed from the columns of $\mathbf{M}$ indexed by S. For $1 \leq p < \infty$ and the vector $\mathbf{x} = [x_1, \ldots, x_m]$, denote $\|\mathbf{x}\|_p = (\sum_{i=1}^m |x_i|^p)^{1/p}$. Instead, $\|\mathbf{u}\|_\infty = \max_l |u_l|$. The $\ell_{a,b}$-mixed-norm, with $a, b \geq 0$, of $\mathbf{z} = [\mathbf{z}_1^\mathsf{T}, \ldots, \mathbf{z}_r^\mathsf{T}]^\mathsf{T}$ is given by $\|\mathbf{z}\|_{a,b}^b = \sum_{j=1}^r \|\mathbf{z}_j\|_a^b$. By convention, $\|\mathbf{z}\|_{a,0} \triangleq \sum_{j=1}^r I(\|\mathbf{z}_j\|_a \neq 0)$, where $I(\cdot)$ is the indicator function, counts the number of non-zero vectors.

II. Problem Setup and Preliminaries

For a sampled system, we obtain a linear relation between measurements and the initial state and forced inputs. We then formulate a group LASSO optimization problem to estimate the initial state and inputs, and to locate the unknown sources.

A. Linear dynamics under sparse forced inputs

Consider the following continuous-time linear system subjected to external inputs:

$$\dot{\mathbf{x}}_c(t) = \mathbf{A}_c \mathbf{x}_c(t) + \mathbf{B}_c \mathbf{u}_c^*(t), \quad t \in \mathbb{R}, \tag{2}$$

where $\mathbf{x}_c(t) \in \mathbb{R}^n$ and $\mathbf{u}_c^*(t) \in \mathbb{R}^m$ is the state and input. We assume the input to be sparse, that is $\|\mathbf{u}_c^*(t)\|_0 \leq m^* << m$ for all $t \in \mathbb{R}$. In the context of power systems, the state $\mathbf{x}_c(t)$ consists of the dynamical states of generators and their control systems, including rotor angles, speed deviations, field excitation voltage, etc. Instead, $\mathbf{u}_c^*(t) = [u_{c,1}^*(t), \ldots, u_{c,m}^*(t)]^\mathsf{T}$ is the vector of inputs triggered by the sources of FOs, among which only m^* locations are active. However, our model in (2), except for sparsity constraints, is general and allows for multi-dimensional un-modeled exogenous stochastic or deterministic disturbances, benign faults, or adversarial attacks.

We consider the discrete-time dynamics of (2) together with a measurement equation:

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}^*[k], \tag{3}$$

$$\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{v}[k], \quad k = 0, 1, \ldots, \tag{4}$$

where $\mathbf{A} = e^{\mathbf{A}_c \delta t}$, $\mathbf{B} = (\int_0^{\delta t} e^{\mathbf{A}_c \tau} d\tau)\mathbf{B}_c$, and δt is the sampling time period, and $\mathbf{u}^*[k] = [u_1[k], \ldots, u_m[k]]^\mathsf{T}$. Further, $\mathbf{y}[k] = [y_1[k], \ldots, y_p[k]]^\mathsf{T} \in \mathbb{R}^p$ is the measurement,

$\mathbf{v}[k] \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is noise, and $\mathbf{C} \in \mathbb{R}^{p\times n}$ is the sensor matrix. In Section IV, we consider dynamics in (3) with process noise, and also relax the diagonal covariance assumption on $\mathbf{v}[k]$.

Let $S = \{j : u_j[k] \neq 0 \text{ for at least one } k \geq 0\} \subset \{1, \ldots, m\}$ and $S^c = \{1, \ldots, m\} \setminus S$. We refer S and S^c to as the active and inactive set. Partition $\mathbf{B}$ as $\mathbf{B} = [\mathbf{B}_S \, \mathbf{B}_{S^c}]$ and $\mathbf{u}[k] = [\mathbf{u}_S^{\mathsf{T}}[k] \, \mathbf{u}_{S^c}^{\mathsf{T}}[k]]^{\mathsf{T}}$, with $\mathbf{u}_{S^c}^*[k] = [u_{i_1}^*[k], \ldots, u_{i_r}^*[k]]$ and $\mathbf{B}_{S^c} = [\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_r}]$, where $i_r \in S^c$ and $r = |S^c| = m - m^*$. Similarly, define $\mathbf{u}_S^*[k]$ and $\mathbf{B}_S$. Then, the input term in (3) can be written as

$$\begin{aligned}\mathbf{B}\mathbf{u}^*[k] = \sum_{j=1}^{m} \mathbf{b}_j u_j^*[k] &= \sum_{j\in S} \mathbf{b}_j u_j^*[k] + \sum_{j \in S^c} \mathbf{b}_j u_j^*[k] \\ &= \mathbf{B}_S \mathbf{u}_S^*[k] + \mathbf{B}_{S^c}\mathbf{u}_{S^c}^*[k].\end{aligned} \tag{5}$$

The above representations will play a key role in formulating our group LASSO problem in Section II-B.

Using (3)-(4), we express the batch measurements $\mathbf{y}$ (see below) as a linear model with added noise. Define the vectors

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}[0] \\ \vdots \\ \mathbf{y}[N] \end{bmatrix}, \mathbf{v} = \begin{bmatrix} \mathbf{v}[0] \\ \vdots \\ \mathbf{v}[N] \end{bmatrix}, \text{ and } \mathbf{u}_j^* = \begin{bmatrix} u_j^*[0] \\ \vdots \\ u_j^*[N] \end{bmatrix}, \tag{6}$$

where $\mathbf{y}, \mathbf{v} \in \mathbb{R}^{p(N+1)}$ and $\mathbf{u}_j^* \in \mathbb{R}^{N+1}$, for all $j \in S \cup S^c$. Here, $N+1$, with $N > 0$ is the length of the estimation horizon. We also define the observability matrix $\mathbf{O} \in \mathbb{R}^{p(N+1)\times n}$ and the impulse response matrix $\mathbf{J}_j \in \mathbb{R}^{p(N+1)\times N+1}$ as

$$\mathbf{O} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^N \end{bmatrix}; \mathbf{J}_j = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{H}_1^{(j)} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{H}_2^{(j)} & \mathbf{H}_1^{(j)} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{H}_N^{(j)} & \mathbf{H}_{N-1}^{(j)} & \ldots & \mathbf{H}_1^{(j)} & \mathbf{0} \end{bmatrix}, \tag{7}$$

where $j \in S \cup S^c$, and, for any $l \geq 1$, the l-th impulse response parameter $\mathbf{H}_l^{(j)} \in \mathbb{R}^{p\times 1}$ is defined as $\mathbf{H}_l^{(j)} := \mathbf{CA}^{l-1}\mathbf{b}_j$.

Let $\mathbf{x}[0] = \mathbf{x}_0^*$ be the unknown initial state. From (3)-(4) and the fact that $\mathbf{Bu}^*[k] = \sum_{j=1}^m \mathbf{b}_j u_j^*[k]$, we observe that

$$\mathbf{y} = \mathbf{O}\mathbf{x}_0^* + \sum_{j=1}^{m} \mathbf{J}_j \mathbf{u}_j^* + \mathbf{v}, \tag{8}$$

where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{p(N+1)})$, $\mathbf{u}_j^*$ is in (6) and $\mathbf{J}_j$ is in (7).

B. Initial State and Unknown Input Estimation under Sparsity Constraints: A Group LASSO for Approach

Based the measurement model in (8), we introduce the group LASSO estimator to estimate $(\mathbf{x}_0^*, \mathbf{u}_1^*, \ldots, \mathbf{u}_m^*)$ and also the active set S. Let $\mathbf{J} = [\mathbf{J}_1^{\mathsf{T}}, \ldots, \mathbf{J}_m^{\mathsf{T}}]$ and $\mathbf{u} = [\mathbf{u}_1^{\mathsf{T}}, \ldots, \mathbf{u}_m^{\mathsf{T}}]$, where $\mathbf{u}_j \in \mathbb{R}^{N+1}$. Recall the definition of $\ell_{p,0}$-norm from the notation section, and consider

$$\begin{bmatrix} \widehat{\mathbf{x}}_0 \\ \widehat{\mathbf{u}} \end{bmatrix} = \underset{\mathbf{x}_0, \mathbf{u}}{\arg\min} \left\{ \frac{1}{2T} \|\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}\mathbf{u}\|_2^2 + \lambda_T \|\mathbf{u}\|_{p,0} \right\}, \tag{9}$$

where the regularization parameter $\lambda_T \geq 0$ and $T = p(N+1)$ is the dimension of $\mathbf{y}$ in (8). The above problem is called subset (or block-column) selection problem because the optimization problem amounts to finding $\mathbf{J}_j$ that contributes to $\mathbf{y}$ in (8).

Unfortunately, (9) is a combinatorial optimization problem and its computationally complexity is exponential in m. We circumvent this difficulty by replacing the $\|\mathbf{u}\|_{p,0}$ with the $\|\mathbf{u}\|_{p,1}$-norm. This is a common relaxation technique widely used in the literature of compressed sensing and statistics; see [25], [26]. Thus, we end up with the group LASSO problem:

$$\begin{bmatrix} \widehat{\mathbf{x}}_0 \\ \widehat{\mathbf{u}} \end{bmatrix} \in \underset{\mathbf{x}_0, \mathbf{u}}{\arg\min} \left\{ \frac{1}{2T} \|\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \mathbf{J}\mathbf{u}\|_2^2 + \lambda_T \|\mathbf{u}\|_{p,1} \right\}. \tag{10}$$

For definiteness, we set $p = 2$, although our analysis extends to the case $p \neq 2$. In the literature, $\|\mathbf{u}\|_{2,1} = \sum_{j=1}^m \|\mathbf{u}_j\|_2$ is referred to as the block or group norm. Our optimization problem in (10) differs from the traditional group LASSO [7] because we do not penalize $\mathbf{x}_0$. This is subtle yet important distinction because in many applications, including power systems, initial state is rarely sparse. In Section VI, we provide details on how to numerically solve (10). Instead, in Section III, for a specific range of λ_T, we show that the group-norm based regularizer promotes group sparsity in $\widehat{\mathbf{u}}$ and that $\widehat{S} = S$ holds with high probability, where $\widehat{S} \triangleq \{j : \widehat{\mathbf{u}}_j \neq 0\}$.

Due to the presence of additive noise in the measurement vector $\mathbf{y}$ in (8), neither the estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}})$ in (10) need to identically match $\boldsymbol{\beta}^* = (\mathbf{x}_0^*, \mathbf{u}^*)$ nor does $\widehat{S} = S$. Thus, we evaluate the quality of our estimates (i.e., the hatted quantities) in a probabilistic sense using the error metrics:

- $\widehat{\boldsymbol{\beta}}$ is said to be ℓ_2-consistent if $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq o(T)$ with probability at least $1 - c_1 \exp(-c_2 T)$, for some $c_1, c_2 > 0$.
- $\widehat{\mathbf{u}}$ is said to be location recovery consistent if $\widehat{S} = S$ with probability at least $1 - c_3 \exp(-c_4 T)$, for $c_3, c_4 > 0$.

Here $o(T)$ implies that the upper bound on the error tends to zero as $T \to \infty$. The ℓ_2-error bound ensures that the estimate $\widehat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}^*$ by increasing $T = p(N+1)$. Instead, the location selection consistency ensures that as as long as T is sufficiently large, $\widehat{S}$ correctly identifies the true sources of FOs.

III. Delayed Estimation and Invariant Zeros

In this section we cull recent results on the initial state and delayed input recovery using finite number of measurements [27], by assuming the knowledge set S. These results provide a starting point to prove our main results in Section IV.

We begin by expressing $\mathbf{y}$ in (8) in a slightly different way. From (5), we have $\mathbf{Bu}^*[k] = \mathbf{B}_S \mathbf{u}_S^*[k] + \sum_{j\in S^c} \mathbf{b}_j u_j^*[k]$. Substituting this fact in (3) and recursively expanding $\mathbf{y}[k]$ in (4) yields us the following model for $\mathbf{y}$ defined in (6).

$$\mathbf{y} = \underbrace{\begin{bmatrix} \mathbf{O} & \mathbf{J}_S \end{bmatrix}}_{\triangleq \boldsymbol{\Psi}_S} \underbrace{\begin{bmatrix} \mathbf{x}_0^* \\ \mathbf{u}_S^* \end{bmatrix}}_{\triangleq \boldsymbol{\beta}_S^*} + \sum_{j\in S^c} \mathbf{J}_j \mathbf{u}_j^* + \mathbf{v}, \tag{11}$$

where $\mathbf{u}_S^*$ and $\mathbf{J}_S \in \mathbb{R}^{p(N+1)\times m^*(N+1)}$ are defined as

$$\mathbf{u}_S^* = \begin{bmatrix} \mathbf{u}_S^*[0] \\ \mathbf{u}_S^*[1] \\ \mathbf{u}_S^*[2] \\ \vdots \\ \mathbf{u}_S^*[N] \end{bmatrix}; \mathbf{J}_S = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{H}_1^{(S)} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{H}_2^{(S)} & \mathbf{H}_1^{(S)} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{H}_N^{(S)} & \mathbf{H}_{N-1}^{(S)} & \ldots & \mathbf{H}_1^{(S)} & \mathbf{0} \end{bmatrix}, \tag{12}$$

with $\mathbf{H}_l^{(S)} = \mathbf{C}\mathbf{A}^{l-1}\mathbf{B}_S$, for all $l \geq 1$. Note that $\mathbf{y}$ in (8) and equals $\mathbf{y}$ in (11). Importantly, $\mathbf{u}_S^*$ in (12) is a concatenation of inputs $\mathbf{u}_S^*[k]$ associated with S from $k = 0$ (top) to N (bottom), but not a concatenation of $\mathbf{u}_j^*$ in (6), for all $j \in S$.

To show that the group LASSO is location recovery consistent, or $\widehat{S} = S$ holds with high probability, $\mathbf{\Psi}_S = \begin{bmatrix}\mathbf{O} & \mathbf{J}_S\end{bmatrix}$ in (11) should be of full column rank. To see this, suppose that $\sigma^2 \approx 0$ and that we know S. Then, by substituting $\mathbf{u}_j^* = \mathbf{0}$, for all $j \in S^c$, and $\mathbf{v} = \mathbf{0}$ in $\mathbf{y}$ in (11), it follows that

$$\mathbf{y} = \mathbf{\Psi}_S \boldsymbol{\beta}_S^*. \tag{13}$$

Thus for a rank deficient $\mathbf{\Psi}_S$, we cannot perfectly recover $\boldsymbol{\beta}_S^* = (\mathbf{x}_0^*, \mathbf{u}_S^*[0], \ldots, \mathbf{u}_S^*[N])$ even with noise-free measurements and with the knowledge of S. However, unfortunately, unlike the model matrices, such as random design and Fourier basis matrices, considered in signal processing and statistics applications, $\mathbf{\Psi}_S$ could be rank deficient. This is so because system in (3)-(4) may not be initial state and input observable [9]; that is, either $\mathbf{O}$ or $\mathbf{J}_S$ is rank deficient, or both $\mathbf{O}$ and $\mathbf{J}_S$ have full ranks, but $[\mathbf{O}\ \mathbf{J}_S]$ is rank deficient.

From the foregoing discussion, it is clear that recovering $\boldsymbol{\beta}_S^*$ and full rank of $\mathbf{\Psi}_S$ are intimately connected. Interestingly, for d-delay invertible linear systems, even when $\boldsymbol{\beta}_S^*$ is not recoverable, a portion of it is perfectly recoverable [9], [27]. In fact, we can recover $\boldsymbol{\beta}_{S,[0:N-d]}^* = (\mathbf{x}_0^*, \mathbf{u}_S^*[0], \ldots, \mathbf{u}_{S,[N-d]})$, where $N \geq d$, from $\mathbf{y}^\mathsf{T} = [\mathbf{y}^\mathsf{T}[0], \ldots, \mathbf{y}^\mathsf{T}[N]]$ Here, $d \geq 0$ is called *delay* and we refer $\boldsymbol{\beta}_{S,[0:N-d]}^*$ to as the delayed input. As a result, we show that a specific sub-matrix of $\mathbf{\Psi}_S$ has full column rank even when $\mathbf{\Psi}_S$ is rank deficient.

We formalize the notion of d-delay. Let $\mathbf{x}_0 = \mathbf{0}$ to note that $\mathbf{\Psi}_S = \mathbf{J}_S$ and $\boldsymbol{\beta}_S^* = \mathbf{u}_S^*$. Substituting $\mathbf{J}_S$ (12) in (13), yields

$$\underbrace{\begin{bmatrix}\mathbf{y}[0]\\ \mathbf{y}[1]\\ \vdots\\ \mathbf{y}[N]\end{bmatrix}}_{\mathbf{y}_N} = \underbrace{\left[\begin{array}{c|ccc}\mathbf{0} & \mathbf{0} & \ldots & \mathbf{0}\\ \hline \mathbf{H}_1^{(S)} & \mathbf{0} & \ldots & \mathbf{0}\\ \vdots & \ddots & \ddots & \vdots\\ \mathbf{H}_N^{(S)} & \mathbf{H}_{N-1}^{(S)} & \ldots & \mathbf{0}\end{array}\right]}_{\triangleq \mathbf{J}_{S,[N:0]}} \underbrace{\left[\begin{array}{c}\mathbf{u}_S^*[0]\\ \hline \mathbf{u}_S^*[1]\\ \vdots\\ \mathbf{u}_S^*[N]\end{array}\right]}_{\mathbf{u}_{S,[0:N]}^*}. \tag{14}$$

Notice that $\mathbf{J}_S = \mathbf{J}_{S,[N:0]}$ and $\mathbf{u}_S^* = \mathbf{u}_{S,[0:N]}^*$. Define

$$\mathbf{J}_{S,[N:0]} = \left[\begin{array}{c|c|cc}\mathbf{M}_N^{(S)} & \mathbf{M}_{N-1}^{(S)} & \ldots & \mathbf{M}_0^{(S)}\end{array}\right], \tag{15}$$

where $\mathbf{M}_l^{(S)}$ denotes the l^{th} block column of $\mathbf{J}_{S,[N:0]}$ labeled right ($l = 0$) to left ($l = N$). By construction $\mathbf{J}_{S,[N:0]}$ is rank deficient because of the zero diagonal blocks in in (14). Thus, we cannot recover $\mathbf{u}_S^*[N]$ using $\mathbf{y}_N$. Further, in several practical applications, $\mathbf{H}_1^{(S)} = \mathbf{C}\mathbf{B}_S = \mathbf{0}$ (or has non-full column rank). This is because sensors may not be located at the inputs. For e.g., in power systems, bus level PMUs do not directly measure power system stabilizer's output. Thus it is impossible to recover $\mathbf{u}_S^*[N-1]$ using $\mathbf{y}_N$.

Definition 1. ***(System delay)*** *For a non-negative integer $d \geq 0$, let $\mathbf{J}_{S,[d:0]}$ be as in* (14)*. System in* (3)-(4)*, with $\mathbf{x}_0^* = \mathbf{0}$, $\mathbf{u}_j^* = \mathbf{0}$, for $j \in S^c$, and $\sigma^2 = 0$, is d-delay left invertible if*

$$\mathrm{Rank}(\mathbf{J}_{S,[d:0]}) - \mathrm{Rank}(\mathbf{J}_{S,[d-1:0]}) = m^*, \tag{16}$$

for $\mathbf{J}_{S,[d:0]}$ defined in (14) *and m^* is the dimension of $\mathbf{u}_S^*[k]$. The smallest d that satisfies* (16) *is denoted as η_S.* □

Throughout we assume $d = \eta_S$ and we set $d \triangleq \infty$ if (16) does not hold for any $d \geq 0$. Suppose that $\eta_S < \infty$. Then, from the rank properties of partitioned matrices [27], $\mathbf{M}_d^{(S)}$ in (15) has full column rank. Thus, there exists a matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{y}_d = \mathbf{u}_S^*[0]$, where $\mathbf{y}_d = [\mathbf{y}^\mathsf{T}[0] \ldots \mathbf{y}^\mathsf{T}[d]]^\mathsf{T}$. We may recover $\mathbf{u}_S^*[1]$ using the residual $\widehat{\mathbf{y}}_{d+1} \triangleq \mathbf{y}_{d+1} - \mathbf{M}_{d+1}^{(S)}\mathbf{u}_S^*[0]$. In fact, $\mathbf{Q}\widehat{\mathbf{y}}_{d+1} = \mathbf{u}_S^*[1]$. By iterating this procedure, we can obtain $\mathbf{u}_{S,[0:N-d]}^* \triangleq [(\mathbf{u}_S^*[0])^\mathsf{T}, \ldots, (\mathbf{u}_S^*[N-d])^\mathsf{T}]^\mathsf{T}$ using $\mathbf{y}_N$.

We relax $\mathbf{x}_0^* = \mathbf{0}$ assumption and extend the rank condition in (16) to recover jointly $\boldsymbol{\beta}_{S,[0:N-d]}^* = (\mathbf{x}_0^*, \mathbf{u}_{S,[0:N-d]}^*)$, as a whole rather than sequentially, using $\mathbf{y}_N$. First, we define the smallest delay for recovering $\mathbf{x}_0^*$ in the presence of input:

$$\mu_S \triangleq \min\{d \geq 0 : \mathrm{Rank}([\mathbf{O}_d\ \mathbf{J}_{S,[d:0]}]) - \mathrm{Rank}(\mathbf{J}_{S,[d:0]}) = n\}, \tag{17}$$

where $\mathbf{O}_d = [\mathbf{C}^\mathsf{T}\ (\mathbf{C}\mathbf{A})^\mathsf{T}\ \ldots,\ (\mathbf{C}\mathbf{A}^d)^\mathsf{T}]$, and n is the dimension of $\mathbf{A}$. The rank condition in (17) says that $\mathbf{O}_d$ has full column rank ($= n$) and that the columns in $\mathbf{O}_d$ are linearly independent of columns in $\mathbf{J}_{S,[d:0]}$. This condition is stronger than system in (3)-(4) being observable, as shown below:

Example 1. *Let $\mathbf{A} = \begin{bmatrix}1 & 2; 0 & 3\end{bmatrix}$, $\mathbf{B}_S = \begin{bmatrix}2 & 3\end{bmatrix}^\mathsf{T}$, and $\mathbf{C} = \begin{bmatrix}1 & 0\end{bmatrix}$. Then $\eta_S = 1$ and $\mathrm{Rank}\,\mathbf{O}_l = 2$, for any $\ell \geq 2$; that is, the system is observable. However, $\mu_S = \infty$. To see this note that the second column of $\mathbf{A}$ is identical to $\mathbf{B}_S$; thus, the matrices $\mathbf{O}_d$ and $\mathbf{J}_{S,[d:0]}$ have some columns in common, and consequently,* (17) *does not hold for any $d \geq 0$.* □

Let $\mathbf{M}_l^{(S)}$ be defined as in (15). For $N \geq d \geq 0$, define $\mathbf{\Psi}_S = \begin{bmatrix}\mathbf{O} & \mathbf{J}_S\end{bmatrix}$; $\mathbf{\Psi}_{S,[N:d]} = [\mathbf{O}\ \mathbf{M}_N^{(S)}\ \ldots\ \mathbf{M}_d^{(S)}]$; and finally, $\mathbf{\Psi}_{S,[d-1:0]} = [\mathbf{M}_{d-1}^{(S)}\ \ldots\ \mathbf{M}_0^{(S)}]$. Note that

$$\mathbf{\Psi}_S = [\mathbf{\Psi}_{S,[N:d]}\ \mathbf{\Psi}_{S,[d-1:0]}] \tag{18}$$

with an understanding that $\mathbf{\Psi}_S = \mathbf{\Psi}_{S,[N:d]}$ for $d = 0$. Let $\mathbf{\Psi}_S^\dagger$ be the pseudo inverse of $\mathbf{\Psi}_S$. The proposition below establishes conditions under which we can recover $(\mathbf{x}_0^*, \mathbf{u}_{S,[0:N-d]}^*)$.

Proposition 1. *Suppose that η_S in* (16) *and μ_S in* (17) *are finite. Then, for $N \geq \max\{\eta_S, \mu_S\}$ with $d \geq \eta_S$, we have*

1) *$\mathbf{\Psi}_{S,[N:d]}$ defined in* (18) *has full column rank.*
2) *$\mathcal{R}(\mathbf{\Psi}_{S,[N:d]}) \cap \mathcal{R}(\mathbf{\Psi}_{S,[d-1:0]}) = \{0\}$.*

Moreover, for $t_S \triangleq (N - d + 1)m^*$ and $m^* = |S|$, we have

$$\begin{bmatrix}\mathbf{x}_0^*\\ \widetilde{\mathbf{u}}_{S,[0:N-d]}^*\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{I}_{n+t_S} & \mathbf{0}_{(n+t_S)\times dm^*}\end{bmatrix}}_{\widetilde{\mathbf{\Pi}}_{S,[0:N-d]}} \mathbf{\Psi}_S^\dagger \mathbf{y}. \tag{19}$$

The proof of this fact is given in [27, Theorem 7]. Part (1) of proposition states that the sub-matrix $\mathbf{\Psi}_{S,[N:d]}$ has full rank even when $\mathbf{\Psi}_{S,[N:0]}$ is rank deficient. This fact plays a vital role in the performance analysis of the group LASSO estimate.

For Proposition 1 to hold, we require $\eta_S, \mu_S < \infty$. Using the notion of zeros and rank of the system matrix (see below), we state verifiable conditions to check if $\eta_S, \mu_S < \infty$. For all $z \in \mathbb{C}$, define the Rosenbrock system matrix:

$$\mathcal{Z}_S[z] \triangleq \begin{bmatrix} z\mathbf{I}_n - \mathbf{A} & -\mathbf{B}_S\\ \mathbf{C} & \mathbf{0}\end{bmatrix}. \tag{20}$$

Let $\mathrm{nRank}\mathcal{Z}_S \triangleq \max_{z\in\mathbb{C}} \mathrm{Rank}\, \mathcal{Z}_S[z]$ be its normal rank. A number $z_0 \in \mathbb{C}$ is called the invariant zero of $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ if $\mathrm{Rank}\, \mathcal{Z}_S[z_0] < \mathrm{nRank}\mathcal{Z}_S$. If $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ has invariant zeros, there exists $\mathbf{u}^*_S \neq 0$ and $\mathbf{x}_0 \neq 0$ such that (noise-free) $y[k] = \mathbf{0}$, for all $k \geq 0$ [28]. (Thus, we cannot distinguish between non-zero and zero inputs from $\mathbf{y}_N$ alone.) Hence, $\eta_S, \mu_S = \infty$.

***Lemma* 2.** *Suppose that* $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ *has no invariant zeros. Then (i)* $\mu_S < \infty$*, and for* $N \geq \mu_S$*, system in* (3)-(4) *is initial state observable; and (ii)* $\eta_S < \infty$ *if* $\mathrm{nRank}\mathcal{Z}_S = n + m^*$.

A proof for the statement (i) can be found in [27, Proposition 5]. Instead, the statement (ii) follows from [9, Theorem 1, pp. 227]. Thus, if $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ satisfies conditions in Lemma 2, the assumptions in Proposition 1 hold. Hence, the sub-matrix $\mathbf{\Psi}_{S,[N:d]}$ has full rank and we can recover $(\mathbf{x}^*_0, \mathbf{u}^*_{S:[0:N-d]})$.

IV. Location Recovery and Estimation Consistency of the group LASSO Estimator

We theoretically investigate the performance of the group LASSO estimator in (10) using the previously stated results for the delayed input estimation. Our results generalize the existing group LASSO's guarantees for static (or non-dynamical) systems [6], [29] to the dynamical systems with delay $d \geq 0$.

Recall that the estimate $\widehat{\boldsymbol{\beta}}$ in (10) is $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_m)$ with $\widehat{\mathbf{u}}_j = [\hat{u}_j[0], \ldots, \hat{u}_j[N]]^\mathsf{T}$. For any $S \subset \{1, \ldots, m\}$, we define $\widehat{\mathbf{u}}_S[k] = [\widehat{u}_{s_1}[k], \ldots, \widehat{u}_{s_{|S|}}[k]]$, for all $k \geq 0$ and $s_j \in S$; that is, we group the estimated inputs associated with the set S. Define $\widehat{\mathbf{u}}^\mathsf{T}_S = [\widehat{\mathbf{u}}^\mathsf{T}_S[0], \ldots, \widehat{\mathbf{u}}^\mathsf{T}_S[N]]$ and $\widehat{\boldsymbol{\beta}}_S = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_S)$. For $S = \{j : \mathbf{u}^*_j \neq 0\}$ and $\widehat{S} = \{j : \widehat{\mathbf{u}}_j \neq 0\}$, we derive conditions under which (i) $\widehat{S} = S$ and (ii) $\|\boldsymbol{\beta}^*_{S,[0:N-d]} - \widehat{\boldsymbol{\beta}}_{S,[0:N-d]}\|_2 \leq \epsilon$, for any $\epsilon > 0$, hold with high probability.

Assumption 3. ***(Identifiability and mutual incoherence conditions [26])*** *Consider the following conditions:*

(A1) Group normalization: There exists a constant $C > 0$ *such that* $\mathbf{O}$ *and* $\mathbf{J}_i$ *in* (7) *satisfy the normalization condition:*

$$\max\{\|\mathbf{O}\|_2, \|\mathbf{J}_1\|_2, \ldots, \|\mathbf{J}_m\|_2\} \leq C\sqrt{T} < \infty. \quad (21)$$

(A2) Model indentifiability: The parameters η_S *in* (16) *and* μ_S *in* (17) *are finite; and* $N \geq \max\{\eta_S, \mu_S\}$ *with* $d \geq \eta_S$*. Further, there exists a constant* $c_{min} > 0$ *such that*

$$\left\| \left(\frac{\mathbf{\Psi}^\mathsf{T}_{S,[N:d]} \mathbf{\Theta} \mathbf{\Psi}_{S,[N:d]}}{T} \right)^\dagger \right\|_2 \leq \frac{1}{c_{min}} < \infty, \quad (22)$$

where $\mathbf{\Psi}_{S,[N:d]}$ *and* $\mathbf{\Psi}_{S,[d-1:0]}$ *are given in Eq (18), and* $\mathbf{\Theta} \triangleq [\mathbf{I} - \mathbf{\Psi}_{S,[d-1:0]} \mathbf{\Psi}^\dagger_{S,[d-1:0]}]$.

(A3) Mutual incoherence: There exists some $\alpha \in [0, 1)$*, referred to as ”mutual incoherence” parameter, such that*

$$\mathrm{MIC} \triangleq \max_{j\in S^c} \left\| \mathbf{J}^\mathsf{T}_j \mathbf{\Psi}_S (\mathbf{\Psi}^\mathsf{T}_S \mathbf{\Psi}_S)^\dagger \right\|_2 \leq \alpha/m^*. \quad (23)$$

Assumptions (A1) and the bound in (22) hold for stable and asymptotically unstable systems[2] if $N < \infty$. However, these assumptions hold only for stable systems if $N \to \infty$. Further, as discussed in Section III, we need the requirement on N, μ_S and η_S in Assumption (A2) to ensure that the initial state and the input associated with the source set S are identifiable. In light of Lemma 2, this requirement is satisfied if $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ has no invariant zeros and that $\mathrm{nRank}\mathcal{Z}_S = n + m^*$. Finally, the constants C and c_{min} do not depend on the horizon N. They capture the inherent complexity in estimating the unknown parameters and play a vital role in the true support recovery.

Assumption (A3) is satisfied if $\mathbf{\Psi}_S$ and $\mathbf{J}_j$ are orthogonal ($\mathbf{J}^\mathsf{T}_j \mathbf{\Psi}_S = \mathbf{0}$, for all $j \in S^c$). Orthogonality is restrictive as number of inputs can be more than outputs, or any column of $\mathbf{B}_S$ in (3) can be a linear combination of $\mathbf{b}_j$, for $j \in S^c$. Nonetheless, (A3) imposes a type of ”approximate” orthogonality between $\mathbf{J}_j$, where $j \in S^c$, and $\mathbf{\Psi}_S$. We quantify this approximation using the parameter α. The ℓ_2-norm bound in (23) could be conservative as the bound depends on m^*. This dependence can be avoided by working with the ℓ_1-norm bound; that is, $\max_{j\in S^c} \left\| \mathbf{J}^\mathsf{T}_j \mathbf{\Psi}_S (\mathbf{\Psi}^\mathsf{T}_S \mathbf{\Psi}_S)^\dagger \right\|_1 \leq \alpha$. However, we stick with (23) as it is useful to derive an upper bound on MIC in (23) using the system transfer function. In simulations, we study the conservatism incurred due to ℓ_2-norm based MIC.

Theorem* 4.** ***(Location recovery consistency) *Suppose that the linear model in* (11) *satisfies assumptions (A1)-(A3) with* $S = \{1, \ldots, m^*\}$*. For some* $\delta > 0$ *and* $c_1 = \log(5)$*, let*

$$\lambda_T = \frac{\sqrt{32}C\sigma}{1-\alpha} \left\{ \sqrt{\frac{(N+1)c_1 + \log(m-m^*)}{T}} + \frac{\delta}{2} \right\}, \quad (24)$$

Then, with probability at least $1 - 4\exp(-T\delta^2/2)$ *we have*

(a) (Non-unique): There are infinitely many solutions of (10).

(b) (No false inclusion): The support set of any estimate $\widehat{\boldsymbol{\beta}}$ *lies in the true support set; that is,* $\widehat{S} \subset S$.

(c) (ℓ_∞ *bounds): The delayed inputs satisfy the bound:* $\max_{j\in S} \|\widehat{\mathbf{u}}_{j,[0:N-d]} - \mathbf{u}^*_{j,[0:N-d]}\|_\infty \leq g_{min}(\lambda_T, \mathbf{\Psi})$*, where*

$$g_{min}(\lambda_T, \mathbf{\Psi}) = \frac{\sigma}{\sqrt{c_{\mathrm{min}}}} \left\{ \sqrt{\frac{2\log((N-d+1)m^*)}{T}} + \delta \right\} + \lambda_T \left\| \mathbf{\Pi}_{S,[0:N-d]} \left(\frac{\mathbf{\Psi}^\mathsf{T}_S \mathbf{\Psi}_S}{T} \right)^\dagger \right\|_\infty, \quad (25)$$

and $\mathbf{\Pi}_{S,[0:N-d]} = [\mathbf{0}_{t_S\times n}\ \mathbf{I}_{t_S}\ \mathbf{0}_{t_S\times dm^*}]$*, where* $t_S = (N-d+1)m^*$*, satisfies* $\mathbf{\Pi}_{S,[0:N-d]}\boldsymbol{\beta}^*_S = \mathbf{u}^*_{S,[0:N-d]}$.

(d) (Minimum input magnitude and no false exclusion): If $\min_{j\in S} \|\mathbf{u}^*_{j,[0:N-d]}\|_\infty \geq g_{min}(\lambda_T, \mathbf{\Psi})$*, we have* $\widehat{S} = S$.

Corollary 5. *Let* $\widehat{\boldsymbol{\beta}}_{S,[0:N-d]} = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_{S,[0:N-d]})$ *and similarly define* $\boldsymbol{\beta}^*_{S,[0:N-d]}$*. Suppose that* $\|(\mathbf{\Psi}^\mathsf{T}_S \mathbf{\Psi}_S/T)^\dagger\|_2 \leq 1/c_{min}$*. Under the assumptions stated in Theorem 4, with probability at least* $1 - \exp(-\delta^2 T/2)$*, we have*

$$\left\| \boldsymbol{\beta}^*_{S,[0:N-d]} - \widehat{\boldsymbol{\beta}}_{S,[0:N-d]} \right\|_2 \leq \frac{2\sigma}{\sqrt{c_{min}}} \left\{ \sqrt{\frac{2c_1(n+t_S)}{T}} + \delta \right\} + \frac{\lambda_T\sqrt{m^*}}{c_{min}}. \quad (26)$$

Proof. See Appendix. □

We use the *primal-dual witness* technique [26], [29], [30] to prove Theorem 4. The details of this technique are in Appendix. The location recovery consistency results in Theorem

[2] At least one of the eigenvalues of $\mathbf{A}$ lie outside the complex unit circle

4 in the literature are referred to as support recovery, here the support means the indices of non-zero $\mathbf{u}_j$ which is S. Below we comment on the scaling laws of Theorem 4.

Part (a) in Theorem 4 states that the group LASSO estimate $\boldsymbol{\beta}$ is non-unique unless the sub-system realized by $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ has zero delay. This is because, for $N > d > 0$, the sub-matrix $\boldsymbol{\Psi}_{S,[N:d]}$ in (18) has full rank, but not $\boldsymbol{\Psi}_S$. However, Part (b) in Theorem 4 states that $\widehat{S} \subseteq S$, for any optimal estimate $\boldsymbol{\beta}$ in (10). Thus, the estimated inputs restricted to the complement set are zero: $\widehat{\mathbf{u}}_{j^c} = \mathbf{0}$, for all $j \in S^c$. Thus, the non-uniqueness of the optimal solution does not effect the location consistency of the group LASSO estimator.

Part (d) in Theorem 4 (d)—a consequence of the ℓ_∞ norm bound in part (b)—says that for $\widehat{S} = S$ to hold (i.e., to detect true inputs correctly) , the true non-zero input signal strength should not be too small, precisely, smaller than β_{min} in (25). The probabilistic result in Theorem 4 also helps determine the number of measurements (N) or sensors (p) required to achieve certain amount of performance. Let us simplify λ_T in (24) to comment on its scaling. By substituting $T = p(N+1)$ and assuming that $\log(m - m^*)/(N+1) \gg c_1$, we have

$$\lambda_T = O\left(\sqrt{\frac{\log(m-m^*)}{p(N+1)}} + \frac{\delta}{2}\right). \tag{27}$$

For $N = 1$, $\lambda_T \approx O(\sqrt{\log(m-m^*)/p})$, which is the optimal λ_T for the standard LASSO problem [26]. Thus, $c_1(N+1)$ in (24) quantifies the number of unknowns in $\mathbf{u}_j^*$, and $N+1$ in T accounts for the number of measurements per sensor.

The choice of λ_T plays an important role in determining if Theorem 4 (c) (that is, $\widehat{S} = S$) holds. In fact, the smaller the λ_T, the smaller the minimum threshold $g_{\min}(\lambda_T, \boldsymbol{\Psi})$. Interestingly, for $\lambda_T = 0$, which happens, say, when $\sigma = 0$, the optimization problem in (10) reduces to the standard ordinary least squares (OLS) problem. Thus, there is no shrinkage of input estimates toward zero. Further, λ_T does not depend on c_{min} in (22) but depends on the group normalization constant C in (21) and the mutual incoherence parameter α in (23).

To study the role of the minimum singular value of $\boldsymbol{\Psi}_S$, denoted by $\rho_{\min}(\boldsymbol{\Psi}_S)$, on $g_{\min}(\lambda_T, \boldsymbol{\Psi})$, we assume that $\boldsymbol{\Psi}_S$ has full rank. Then from the standard norm bounds, we have

$$\kappa_1 + \frac{\lambda_T T\sqrt{\kappa_2}}{\rho^2_{\min}(\boldsymbol{\Psi}_S)} \geq g_{\min}(\lambda_T, \boldsymbol{\Psi}) \geq \kappa_1 + \frac{\lambda_T T}{\sqrt{\kappa_2}\rho^2_{\min}(\boldsymbol{\Psi}_S)},$$

where κ_1 is the first term on the right side of the equality in (25) and $\kappa_2 = (N+1)m^*$ is the dimension of $\mathbf{u}_S^*$. For fixed C (defined in (21)) and κ_2, from the preceding inequality, it is clear that larger $\sigma^2_{\min}(\boldsymbol{\Psi}_S)$ requires smaller $g_{\min}(\lambda_T, \boldsymbol{\Psi})$ because the effective signal strength of $\boldsymbol{\Psi}_S\mathbf{u}_S^*$ is large. Instead, smaller $\sigma^2_{\min}(\boldsymbol{\Psi}_S)$ requires higher $g_{\min}(\lambda_T, \boldsymbol{\Psi})$, requiring $\mathbf{u}_S^*$ to be large. If not, the strength of $\boldsymbol{\Psi}_S\mathbf{u}_S^*$ decreases. Finally, from (24), we observe that λ_T is an increasing function of $\alpha \in [0, 1)$; thus, higher the α larger is the $g_{\min}(\lambda_T, \boldsymbol{\Psi})$. Recall that α is large if $\mathbf{J}_j$, for $j \in S^c$, is highly correlated with $\boldsymbol{\Psi}_S$.

We now comment on the ℓ_2-error bound between $\boldsymbol{\beta}^*_{S,[0:N-d]}$ and $\widehat{\boldsymbol{\beta}}_{S,[0:N-d]}$ given in Corollary 5. First, the bound depends on the number of unknown parameters $n + t_S = n + (N - d + 1)m^*$, i.e., the dimension of the initial state and delayed input. Letting $T = p(N+1) \gg n$, we observe that the first term of the bound in (26) scales as $O((2\sigma/\sqrt{c_{\min}})(\sqrt{m^*/p} + \delta))$. Thus, more sensors result in less error. However, the bound is loose for large values of λ_T. To remedy this shortcoming, we consider the following OLS estimate:

$$\widehat{\boldsymbol{\beta}}^{(OLS)}_{\widehat{S},[0:N-d]} \triangleq \widetilde{\boldsymbol{\Pi}}_{\widehat{S},[0:N-d]}(\boldsymbol{\Psi}^\dagger_{\widehat{S}}\mathbf{y}), \tag{28}$$

where $\widetilde{\boldsymbol{\Pi}}_{\widehat{S},0:N-d}$ is defined similar to $\widetilde{\boldsymbol{\Pi}}_{S,0:N-d}$ in (19). We present the second main result of this section: an oracle bound on the error $\|\boldsymbol{\beta}^*_{S,[0:N-d]} - \widehat{\beta}^{(OLS)}_{\widehat{S},[0:N-d]}\|_2$.

Theorem 6. ***(ℓ_2-consistency: oracle bounds)*** *Suppose that the hypotheses in Theorem 4 hold. Then, for any* $\delta, \delta_1 > 0$, *with probability at least* $1 - 4\exp(-T\delta^2/2) - \delta_1$,

$$\left\|\boldsymbol{\beta}^*_{S,[0:N-d]} - \widehat{\boldsymbol{\beta}}^{(OLS)}_{\widehat{S},[0:N-d]}\right\|_2 \leq \frac{4\sigma}{\sqrt{c_{min}}}\left\{\sqrt{\frac{(n+t_S)}{T}}\right\} + \frac{2\sigma}{\sqrt{c_{min}}}\left\{\sqrt{\frac{1}{T}\log\left(\frac{1}{\delta_1}\right)}\right\}, \tag{29}$$

The proof is in Appendix. Similar to the bound in Corollary 5, the first term in (29) is $O((2\sigma/\sqrt{c_{\min}})\sqrt{m^*/p})$; however, the second term in (29) does not depend on λ_T and it approaches zero as $T \to \infty$. Thus, the overall error is dictated by m^*/p. We call the bound in (29) as the oracle because the bound holds for $\widehat{\boldsymbol{\beta}}^{(OLS)}_{S,[0:N-d]}$, albeit with probability $1 - \delta_1$.

A. Mutual Incoherence: Frequency Domain

Thus far we discussed the location recovery- and estimation-consistency of the group LASSO estimator in (10) assuming that assumptions in (A1)-(A3) hold of which the first two are satisfied by stable dynamical systems with $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ having no invariant zeros[3]. However, (A3) might not hold for arbitrary systems, and moreover, verifying (23) can be computationally demanding when either N (the measurement horizon) or n (dimension of system matrix $\mathbf{A}$) is large. In what follows, we bound $\max_{j\in S^c}\|\mathbf{J}_j^\mathsf{T}\boldsymbol{\Psi}_S(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger\|_2$ in (23) using a quantity that depends on the transfer function matrices associated with $(\mathbf{A}, \mathbf{B}_S, \mathbf{C})$ and $(\mathbf{A}, \mathbf{b}_j, \mathbf{C})$, for $j \in S^c$. The advantage is that this upper bound can computed efficiently, as it depends only on the lower dimensional system matrices but not on N.

To simplify the exposition, we let $\mathbf{x}_0 = \mathbf{0}$; thus, $\boldsymbol{\Psi}_S = \mathbf{J}_S$. Let $\mathcal{G}_S[z] \triangleq \mathbf{C}(z\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}_S$; $\mathcal{G}_{S^c}[z] \triangleq \mathbf{C}(z\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}_{S^c}$; and $\mathcal{G}_j[z] = \mathbf{C}(z\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{b}_j$, where $j \in S^c$ and $\mathbf{B}_{S^c}$ is the matrix composed of columns $\mathbf{b}_j$ with $j \in S^c$.

Theorem 7. *Assumption (A3) holds if* $\mathrm{nRank}\mathcal{Z}_S = n + m^*$ *and*

$$\max_{j\in S^c}\max_{\{z\in\mathbb{C}:|z|=1\}}\left\|\mathcal{G}_S^+[z]\mathcal{G}_j[z]\right\|_2 \leq \alpha/m^* < 1. \tag{30}$$

Proof. See Appendix. □

We refer to the expression in (30) as the frequency domain mutual incoherence condition. To verify Assumption (A3), we need to check if the worst case gain of the matrix $\mathcal{G}_S^+[z]\mathcal{G}_j[z]$ is bounded above by α/m^*; see Fig. 1. If computing (30) is

[3] Systems having invariant zeros lie in a zero measure set [28].

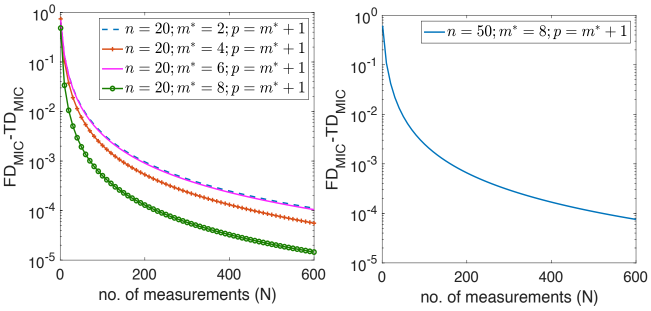

Fig. 1. Illustration of Theorem 7 for systems generated using MATLAB. The number of sources $m = 10$. In both panels, the y-axis, $FD_{\text{MIC}} - TD_{\text{MIC}}$, is the error between frequency- and time-domain MIC. (Left panel) We fix $n = 20$ and plot $FD_{\text{MIC}} - TD_{\text{MIC}}$ for several values of m^*. (Right panel) For a large matrix $\mathbf{A}$, we fix m^* and p, and plot $FD_{\text{MIC}} - TD_{\text{MIC}}$ for several values of n. In both panels, the error is positive and is monotone in N implying that $FD_{\text{MIC}} \geq TD_{\text{MIC}}$, as predicted by Theorem 7.

prohibitive for each $j \in S^c$, we can use the weaker condition: $\max_{\{z\in\mathbb{C}:|z|=1\}} \left\|\mathcal{G}_S^+[z]\mathcal{G}_{S^c}[z]\right\|_2 \leq \alpha/m^* < 1$. To appreciate the condition in (30), take the $\mathcal{Z}$-transform of system in (3)-(4)

$$\mathbf{y}[z] = \mathcal{G}_S[z]\mathbf{u}_S[z] + \sum_{j\in S^c} \mathcal{G}_j[z]\mathbf{u}_j[z], \quad \forall z \notin \text{spec}(\mathbf{A}).$$

By pre-multiplying the above identity with $\mathcal{G}_S^+[z]$, we have $\mathcal{G}_S^+[z]\mathbf{y}[z] = z^{-d}\mathbf{u}_S[z] + \sum_{j\in S^c} \mathcal{G}_S^+[z]\mathcal{G}_j[z]\mathbf{u}_j[z]$, where we used the fact that the $\mathcal{G}_S[z]$ is d delay invertible, and hence, $\mathcal{G}_S^+[z]\mathcal{G}_S[z] = z^{-d}\mathbf{I}$. Thus to recover $\mathbf{u}_S[z]$ accurately, the gain $\|\mathcal{G}_S^+[z]\mathcal{G}_j[z]\|_2$ or $\|\mathcal{G}_S^+[z]\mathcal{G}_{S^c}[z]\|_2$ needs to be small.

We highlight three cases where (30) holds: (i) $\mathcal{R}(\mathcal{G}_{S^c}[z]) \subseteq \mathcal{R}^\perp(\mathcal{G}_S^\dagger[z]) = \mathcal{R}^\perp(\mathcal{G}_S^T[z])$; that is, the columns of $\mathcal{G}_{S^c}[z]$ lie in the left nullspace of $\mathcal{G}_S[z]$; (ii) $\mathcal{G}[z] = [\mathcal{G}_S[z] \ \mathcal{G}_{S^c}[z]]$ is all-pass[4]; and (iii) each column of $\mathcal{G}_{S^c}[z]$ is a scaled column of $\mathcal{G}_S[z]$ for some scaling factor $\alpha \in [0, 1)$. The first two cases are rather strong and do not allow columns of $\mathcal{G}_{S^c}[z]$ to be in the range space of $\mathcal{G}_S[z]$. Instead, (iii) models another extreme where the range spaces of $\mathcal{G}_S[z]$ and $\mathcal{G}_{S^c}[z]$ are aligned with each. The latter case in the compressed sensing literature is referred to as overcomplete dictionaries [31].

Our results extend to System in (3)-(4) driven by process noise (for e.g., in the power system, the noise is load fluctuations); see our extended paper [32, Section IV.A.] for details.

V. Simulations

We illustrate the performance of the group LASSO estimator on a large-scale power network and a random system. The following proposition states that the unknown input and initial state can be estimated in two stages. Consequently, we use off-the-shelf ADMM [10] to estimate the input first and then use this estimate to compute the initial state.

Proposition 8. *Suppose that system in* (3)-(4) *is observable. The optimization problem* (10) *is equivalent to*

$$\widehat{\mathbf{u}} = \underset{\mathbf{u}\in\mathbb{R}^{mT}}{\arg\min} \frac{1}{2T} \|\boldsymbol{R}(\mathbf{y} - \mathbf{J}\mathbf{u})\|_2^2 + \lambda_T \sum_{j=1}^{m} \|\mathbf{u}_j\|_2 \,, \quad (31)$$

$$\widehat{\mathbf{x}}_0 = \mathbf{O}^\dagger(\mathbf{y} - \mathbf{J}\widehat{\mathbf{u}}), \quad (32)$$

where $\mathbf{O}^\dagger = (\mathbf{O}^\mathsf{T}\mathbf{O})^{-1}\mathbf{O}^\mathsf{T}$ *and* $\boldsymbol{R} = \mathbf{I} - \mathbf{O}\mathbf{O}^\dagger$.

The proof follows from the KKT conditions in (33)-(34). The inputs to the ADMM [10] are $(\mathbf{A}, \mathbf{B}, \mathbf{C})$, the measurement $\mathbf{y}$, and $\lambda_T \geq 0$. The two-stage estimation method is one way to implement the group LASSO numerically. One may use other numerical algorithms to estimate $(\mathbf{x}_0^*, \mathbf{u}^*)$ in one shot.

We evaluate the group LASSO estimator's localization performance using the false-positive rate (FPR):= $|S^c \cap \widehat{S}|/|S^c|$, the false-negative rate (FNR):= $|S \cap \widehat{S}^c|/|S|$, and the exact recovery rate (ERR):= $(|S\cap\widehat{S}| + |S^c \cap \widehat{S}^c|)/m$. Thus, the FPR and FNR measure the proportion of inputs falsely identified and left out. Instead, we quantify the estimation performance using the metrics: $\|\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\|_2/\|\mathbf{x}_0^*\|_2$ and $\|\mathbf{u}^* - \widehat{\mathbf{u}}\|_2/\|\mathbf{u}\|_2$. For the test cases below, the results are averaged over 50 runs.

(Power system) We apply our estimator in (31) to localize the sources of forced oscillatory inputs in the IEEE 68 bus system 16 machine system (see Fig. 2). Each machine (or generator) consists of ten states, including rotor angle, speed, and the states of the AVR (automatic voltage regulator) and PSS. We model FOs as inputs injected by the AVRs and use bus voltage magnitudes as measurements. For the sampling time $\delta t = 0.1$, we obtained the system matrices $\mathbf{A} \in \mathbb{R}^{160\times160}$, $\mathbf{B} \in \mathbb{R}^{160\times16}$, and $\mathbf{C} \in \mathbb{R}^{p\times160}$, where $p \leq 68$, using the Power System Toolbox [33]. Among $m = 16$ possible inputs, we assume $m^* = 3$ with the following inputs: $u_1^*[k] = 0.5\sin[(2\pi f\delta t)\,k] + w[k]$, $u_6^*[k] = 0.6\sin[(2\pi f\delta t)\,k] + w[k]$, and $u_{13}^*[k] = 0.7\sin[(2\pi f\delta t)\,k] + w[k]$, where $f = 1.5\,\mathcal{U}(0, 1)$ and $w[k] \sim \mathcal{N}(0, 0.05^2)$. We set $p = 4$ and choose sensor locations arbitrarily with the only exception that these are non-collocated with inputs (shown in Fig. 2). Let $\mathbf{x}_0 = \mathbf{0}$ (the non-zero case is considered in the subsequent case). Finally, we let $N = 100$ and the noise variance $\sigma^2 = 0.01$.

In Fig. 3, we plot the FPR, FNR, and ERR with respect to λ_T. As expected, the FNR increases with λ_T whereas the FPR decreases with λ_T, although not monotonically. From the bottom left panel, we can infer that values of $\lambda_T \in (0.3, 0.4)$ yield maximum ERR. In the bottom right panel, note that for $\lambda_T = 0.288$, the group LASSO estimator accurately localized inputs among 40 out of 50 runs. In Fig. 4, for a measurement realization where the group LASSO estimator identified true locations, we plot the inputs estimated by the group LASSO and the reduced model based OLS estimators.

(Large-scale random system) Following [22], we generate matrices as follows: $\mathbf{A}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/n)$; $\mathbf{C}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$; and $\mathbf{B}^\mathsf{T} = \begin{bmatrix}\mathbf{I}_m^\mathsf{T} & \mathbf{0}^\mathsf{T}\end{bmatrix}$. We let $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and the measurement noise variance parameter $\sigma = 0.01$. We set $n = 50$, $m = 30$, and $m^* = 5$. The active set $S = \{1, 2, 3, 4, 5\}$ and $u_j[k]$ is sampled uniformly on $[-2, 2]$, for all $j \in S$ and $k \in [N]$. The sensors measures the first $p(\leq n)$ states. In Fig. 5, for $p = 15$, we plot the average estimation error metrics as a function of the measurement horizon (N). In both the panels, estimation errors remain uniform across N because the number of (to be estimated) inputs also increase with N. Given the relation in (32), the estimation error of $\mathbf{x}_0^*$ is slightly higher than that of the unknown input. Finally, for greater estimation accuracy, one can always use the reduced model-based OLS estimator.

In Fig. 6 we show the average mutual incoherence in (23)

[4] A real rational transfer function matrix $\mathcal{G}[z]$ is all-pass if $\mathcal{G}[z]\mathcal{G}[1/z] = \mathbf{I}$.

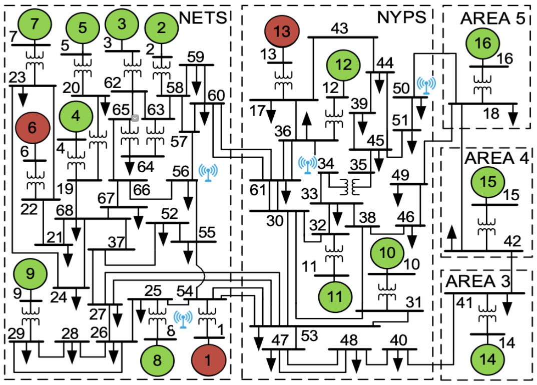

Fig. 2. IEEE 16 machine 68 bus system [34]. Circles, arrows, and curly windings, respectively, denote generator buses, load buses, and transformers. The FO input enters through set points of AVRs associated with the generators at buses $\{1, 6, 13\}$ (red circles). Sensors are located at buses $\{8, 34, 50, 56\}$.

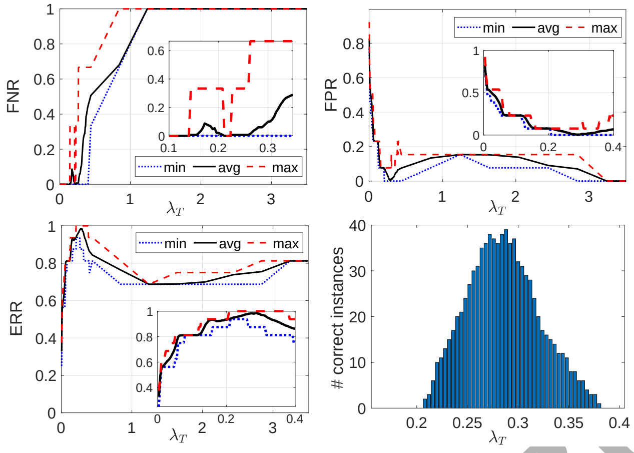

Fig. 3. Group LASSO's performance on IEEE 16 machine 68 bus system: False negative rate (FNR), false positive rate (FPR), exact recovery rate (ERR), and the number of runs out of the 50 independent runs in which the group LASSO exactly recovered the support.

as a function of p for $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{x}_0 \neq \mathbf{0}$. We computed both ℓ_1- and ℓ_2-norm based MICs. As pointed out in Section IV, and confirmed by our plots in the left panel of Fig. 6, ℓ_2-norm based MIC assumption is stronger than the ℓ_1-norm. Further, when $\mathbf{x}_0 = \mathbf{0}$, the MIC is satisfied (that is, less than one) for as few as $p = 6$ sensors. Instead, when $\mathbf{x}_0 \neq \mathbf{0}$, we need at least $p = 18$ sensors to ensure that MIC is less than one.

VI. Conclusion

We have studied a group LASSO estimator for localizing the sparse set of sources of forced inputs and estimating these inputs along with the initial state in d-delay left invertible linear dynamical systems. Under certain natural conditions, we showed that our estimator is well defined, and the underlying estimate is non-unique for $d \geq 0$. However, with high probability, we showed that the support of any optimal estimate recovers the true sparse set if (i) the subsystem associated with the sparse set of inputs has no invariant zeros, and (ii) the observability and impulse response matrices of the overall system satisfy a mutual-incoherence type condition. In doing so, we have extended the existing theory of the group LASSO estimator for static regression models to the models generated by linear dynamical systems. Another key contribution in our paper is that we derived a connection between the time- and frequency-domain mutual incoherence conditions. The former imposes certain restrictions on the column space of the impulse response matrix; instead, the latter does so on the column space of the transfer function matrix. Furthermore, the frequency-domain condition is computationally easier to verify than its time-domain counterpart. Importantly, it pro-

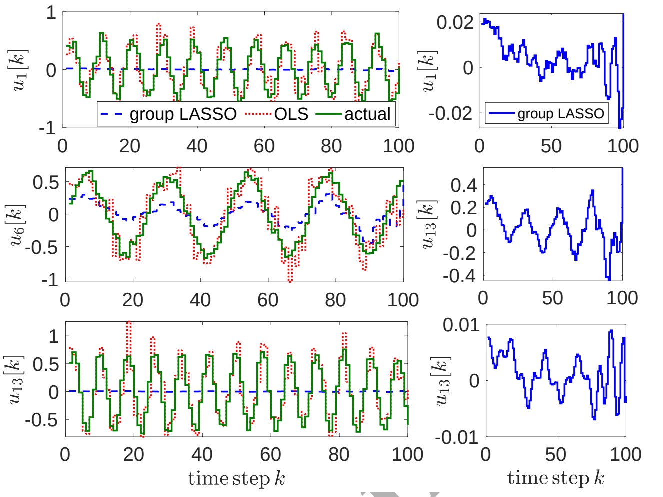

Fig. 4. FO inputs recovered by the group LASSO and OLS estimators. We used (28) to compute the OLS estimate using the locations recovered by the group LASSO. (Left panel) As predicted by Theorem 6, the OLS provides a better estimate than the LASSO estimator. (Right panel) zoomed plot of the group LASSO estimate.

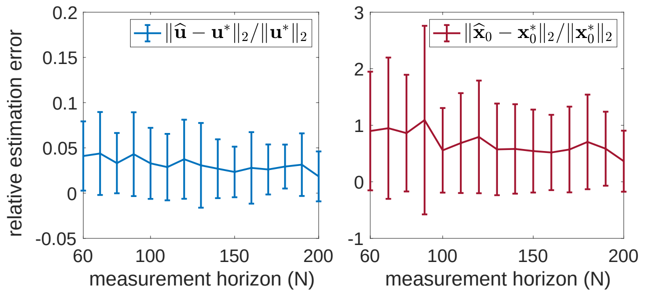

Fig. 5. Estimation error. Left panel: unknown inputs. Right panel: initial state.

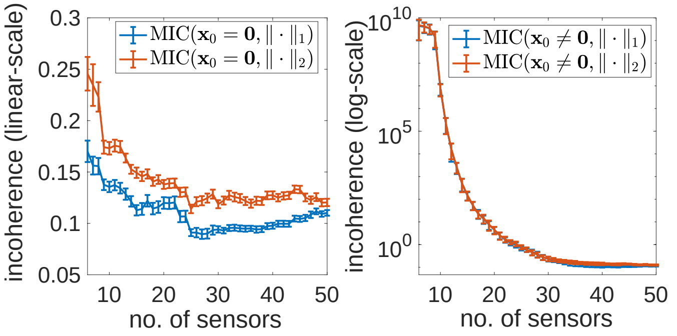

Fig. 6. Mutual incoherence as a function of number of sensors.

vides insight into the structural aspects of transfer matrices associated with the zero and non-zero inputs. Finally, we have validated the performance of the proposed on the IEEE 68-bus, 16-machine power system, and a large-scale synthetic model.

VII. Appendix

A. KKT conditions and PDW Construction

Proposition 9. ***(Karush-Kuhn-Tucker (KKT) conditions)*** *A necessary and sufficient condition for* $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}})$*, with* $\widehat{\mathbf{u}}^\mathsf{T} = [\widehat{\mathbf{u}}_1^\mathsf{T}, \ldots, \widehat{\mathbf{u}}_m^\mathsf{T}]$*, to be a solution of* (10) *is*

$$-\frac{1}{T}\mathbf{O}^\mathsf{T}[\mathbf{y} - \mathbf{O}\widehat{\mathbf{x}}_0 - \sum_{j=1}^{m} \mathbf{J}_j\widehat{\mathbf{u}}_j] = \mathbf{0} \qquad (33)$$

$$-\frac{1}{T}\mathbf{J}_i^\mathsf{T}[\mathbf{y} - \mathbf{O}\widehat{\mathbf{x}}_0 - \sum_{j=1}^{m} \mathbf{J}_j\widehat{\mathbf{u}}_j] + \lambda_T\widehat{\mathbf{z}}_j = \mathbf{0} \qquad (34)$$

for all $j \in \{1, \ldots, m\}$*. Here,* $\widehat{\mathbf{z}}_j$ *is the subgradient of* $\|\widehat{\mathbf{u}}_j\|_2$*; that is,* $\widehat{\mathbf{z}}_j = \widehat{\mathbf{u}}_j/\|\widehat{\mathbf{u}}_j\|_2$ *if* $\widehat{\mathbf{u}}_j \neq \mathbf{0}$*, else* $\widehat{\mathbf{z}}_j \in \{\mathbf{q} : \|\mathbf{q}\|_2 \leq 1\}$*.*

The proof follows by taking the derivative of the objective function in (10) with respect to $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}})$ and using the subgradient characterization of the $\|\cdot\|_2$-norm (see [35, Appendix B]). Without loss of generality let $S = \{1, \ldots, m^*\}$ and $S^c = \{m^* + 1, \ldots, m\}$. Let $\widehat{\mathbf{u}}_S^\mathsf{T}[k] = [\widehat{\mathbf{u}}_1[k] \ldots, \widehat{\mathbf{u}}_{m^*}[k]]$, for all $k \in \{0, \ldots, N\}$, where $\widehat{\mathbf{u}}_j[k]$ is the k-th entry of $\widehat{\mathbf{u}}_j$. Define $\widehat{\mathbf{u}}_S^\mathsf{T} = [\widehat{\mathbf{u}}_S^\mathsf{T}[0], \ldots, \widehat{\mathbf{u}}_S^\mathsf{T}[N]]$. Thus,

$$[\widehat{\mathbf{u}}_1^\mathsf{T}, \ldots, \widehat{\mathbf{u}}_{m^*}^\mathsf{T}]^\mathsf{T} = \mathbf{P}\widehat{\mathbf{u}}_S, \qquad (35)$$

for some permutation matrix $\mathbf{P}$. Further, we can verify that $[\mathbf{J}_1 \ldots, \mathbf{J}_{m^*}]\mathbf{P} = \mathbf{J}_S$ (as in (12)). Let $\widehat{\boldsymbol{\beta}}_S \triangleq [\widehat{\mathbf{x}}_0^\mathsf{T} \; \widehat{\mathbf{u}}_S^\mathsf{T}]^\mathsf{T}$ Then,

$$\mathbf{O}\widehat{\mathbf{x}}_0 + \sum_{j \in S} \mathbf{J}_j\widehat{\mathbf{u}}_j = \mathbf{\Psi}_S\widehat{\boldsymbol{\beta}}_S. \qquad (36)$$

Let $\widetilde{\mathbf{J}}_{S^c} = [\mathbf{J}_{m^*+1}, \ldots, \mathbf{J}_m]$. Then, (33)-(34) can be written as

$$-\frac{1}{T}\begin{bmatrix}\mathbf{\Psi}_S^\mathsf{T} \\ \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\end{bmatrix}[\mathbf{y} - \mathbf{O}\widehat{\mathbf{x}}_0 - \sum_{j=1}^{m} \mathbf{J}_j\widehat{\mathbf{u}}_j] + \lambda_T\left[\begin{array}{c}\mathbf{0} \\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S \\ \hline \widehat{\mathbf{z}}_{S^c}\end{array}\right] = \left[\begin{array}{c}\mathbf{0} \\ \mathbf{0} \\ \hline \mathbf{0}\end{array}\right], \qquad (37)$$

where $\widehat{\mathbf{z}}_S^\mathsf{T} = [\widehat{\mathbf{z}}_1^\mathsf{T}, \ldots, \widehat{\mathbf{z}}_{m^*}^\mathsf{T}]$, and $\widehat{\mathbf{z}}_{S^c}^\mathsf{T} = [\widehat{\mathbf{z}}_{m^*+1}^\mathsf{T}, \ldots, \widehat{\mathbf{z}}_m^\mathsf{T}]$.

Primal-dual witness (PDW) construction: We prove Theorems 6 and 4 using the PDW method [5] in [30]. Upon successful completion, this method returns a pair $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{z}})$ that is primal-dual optimal, and it act as a witness (or certificate) for the fact that the group LASSO estimate has the true support.

(a) Set $\widehat{\mathbf{u}}_j = 0$, for all $j \in S^c$.

(b) Let $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_1 \ldots, \widehat{\mathbf{u}}_{m^*})$ be the solution of the sub-problem:

$$\min_{\substack{\mathbf{x}_0; \\ \mathbf{u}_1, \ldots, \mathbf{u}_{m^*}}} \frac{1}{2T}\left\|\mathbf{y} - \mathbf{O}\mathbf{x}_0 - \sum_{j=1}^{m^*} \mathbf{J}_j\mathbf{u}_j\right\|_2^2 + \lambda_T \sum_{j=1}^{m^*} \|\mathbf{u}_j\|_2. \qquad (38)$$

Choose the sub-gradient $\widehat{\mathbf{z}}_S = [\widehat{\mathbf{z}}_1^\mathsf{T}, \ldots, \widehat{\mathbf{z}}_{m^*}^\mathsf{T}]^\mathsf{T}$ such that

$$-\frac{1}{T}\mathbf{\Psi}_S^\mathsf{T}\left[\mathbf{y} - \mathbf{O}\widehat{\mathbf{x}}_0 - \textstyle\sum_{j=1}^{m^*} \mathbf{J}_j\widehat{\mathbf{u}}_j\right] + \lambda_T\begin{bmatrix}\mathbf{0} \\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix} = \mathbf{0}. \qquad (39)$$

(c) Solve $\widehat{\mathbf{z}}_{S^c} = [\widehat{\mathbf{z}}_{m^*+1}^\mathsf{T}, \ldots, \widehat{\mathbf{z}}_m^\mathsf{T}]^\mathsf{T}$ using (37), and check if $\|\widehat{\mathbf{z}}_j\|_2 \leq 1$, for all $j \in S^c = \{m^* + 1, \ldots, m\}$.

By construction, $(\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_1 \ldots, \widehat{\mathbf{u}}_{m^*})$, $\widehat{\mathbf{z}}_S$, and $\widehat{\mathbf{z}}_{S^c}$ that we determined in steps (a), and (b) satisfy conditions in (37). The PDW construction is said to be successful if $\widehat{\mathbf{z}}_{S^c}$ satisfies the strict dual feasibility condition: $\|\widehat{\mathbf{z}}_j\|_2 \leq 1$, for all $j \in S^c$.

For the estimate in (38), define

$$\widehat{\boldsymbol{\beta}}_{\text{PDW}} = (\widehat{\mathbf{x}}_0, \widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_{m^*}, \underbrace{\mathbf{0}_{(N+1)}, \ldots, \mathbf{0}_{(N+1)}}_{m-m^*}). \qquad (40)$$

Lemma **10.** *Suppose that the PDW construction succeeds. Then,* $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\text{PDW}}$ *is an optimal solution of* (10)*.*

Proof. We adapt the method in [30, Lemma 7.23]. Let $d \geq 0$. Because the PDW construction succeeds, $\widehat{\boldsymbol{\beta}}_{\text{PDW}}$ is an optimal solution of (10) satisfying the conditions in Proposition 9.

Let $\mathbf{u}^\mathsf{T} = [\mathbf{x}_0^\mathsf{T}, \mathbf{u}_1^\mathsf{T}, \ldots, \mathbf{u}_m^\mathsf{T}]$ and $F(\mathbf{u}) \triangleq \frac{1}{2T}\|\mathbf{y} - \mathbf{O}\mathbf{x}_0 + \sum_{j=1}^{m} \mathbf{J}_j\mathbf{u}_j\|_2^2$. Let $\nabla F(\mathbf{u})$ be the gradient of $F(\mathbf{u})$ at $\mathbf{u}$. Then, for any other optimal solution $\widetilde{\mathbf{u}}$ of (10), we must have $F(\widehat{\mathbf{u}}) + \lambda_T\widehat{\mathbf{z}}^\mathsf{T}\widehat{\mathbf{u}} = F(\widetilde{\mathbf{u}}) + \lambda_T \sum_{j=1}^{m} \|\widetilde{\mathbf{u}}_j\|_2$, where $\widehat{\mathbf{z}}^\mathsf{T} = [\widehat{\mathbf{z}}_0^\mathsf{T}, \widehat{\mathbf{z}}_1^\mathsf{T}, \ldots, \widehat{\mathbf{z}}_m^\mathsf{T}]$ is the subgradient and we used the fact that $\sum_{j=1}^{m} \widehat{\mathbf{z}}_j^\mathsf{T}\widehat{\mathbf{u}}_j = \sum_{j=1}^{m} \|\widehat{\mathbf{u}}_j\|_2$. The latter holds which holds because $\widehat{\mathbf{u}}_j = \mathbf{0}$ for $j \in S^c$ and $\widehat{\mathbf{z}}_j = \widehat{\mathbf{u}}_j/\|\widehat{\mathbf{u}}_j\|_2$ for $j \in S$. Thus, $F(\widehat{\mathbf{u}}) - \lambda_T\widehat{\mathbf{z}}^\mathsf{T}(\widetilde{\mathbf{u}} - \widehat{\mathbf{u}}) = F(\widetilde{\mathbf{u}}) + \lambda_T \sum_{j=1}^{m} \|\widetilde{\mathbf{u}}_j\|_2 - \lambda_T\widehat{\mathbf{z}}^\mathsf{T}\widetilde{\mathbf{u}}$. Instead, from (33)-(34), we have $\lambda_T\widehat{\mathbf{z}} = -\nabla F(\widehat{\mathbf{u}})$. Thus,

$$F(\widehat{\mathbf{u}}) + \nabla F(\widehat{\mathbf{u}})^\mathsf{T}(\widetilde{\mathbf{u}} - \widehat{\mathbf{u}}) - F(\widetilde{\mathbf{u}}) = \lambda_T(\sum_{j=1}^{m} \|\widetilde{\mathbf{u}}_j\|_2 - \widehat{\mathbf{z}}^\mathsf{T}\widetilde{\mathbf{u}}).$$

By convexity of F, $F(\widehat{\mathbf{u}}) + \nabla F(\widehat{\mathbf{u}})^\mathsf{T}(\widetilde{\mathbf{u}} - \widehat{\mathbf{u}}) - F(\widetilde{\mathbf{u}}) < 0$. Thus, $\sum_{j=1}^{m} \|\widetilde{\mathbf{u}}_j\|_2 \leq \widehat{\mathbf{z}}^\mathsf{T}\widetilde{\mathbf{u}} = \sum_{j=1}^{m} \widehat{\mathbf{z}}_j^\mathsf{T}\widetilde{\mathbf{u}}_j$, where $\widehat{\mathbf{z}}_0 = \mathbf{0}$. Since $\sum_{j=1}^{m} \widehat{\mathbf{z}}_j^\mathsf{T}\widetilde{\mathbf{u}}_j \leq \sum_{j=1}^{m} \|\widehat{\mathbf{z}}_j\|_2\|\widetilde{\mathbf{u}}_j\|_2 \leq \sum_{j=1}^{m} \|\widetilde{\mathbf{u}}_j\|_2$, we have $\sum_{j=1}^{m} \|\widetilde{\mathbf{u}}_j\|_2 = \sum_{j=1}^{m} \widehat{\mathbf{z}}_j^\mathsf{T}\widetilde{\mathbf{u}}_j$. Further, because $\|\widehat{\mathbf{z}}_j\|_2 < 1$, this equality holds only if $\widetilde{\mathbf{u}}_j = \mathbf{0}$, for all $j \in S^c$. To see this note that $\sum_{j=1}^{m} \widehat{\mathbf{z}}_j^\mathsf{T}\widetilde{\mathbf{u}}_j = \sum_{j \in S} \widehat{\mathbf{z}}_j^\mathsf{T}\widetilde{\mathbf{u}}_j + \sum_{j \in S^c} \|\widehat{\mathbf{z}}_j\|_2\|\widetilde{\mathbf{u}}_j\|_2 \cos(\theta_j)$, where θ_j is the angle between $\widehat{\mathbf{z}}_j$ and $\widetilde{\mathbf{u}}_j$, and $\|\widehat{\mathbf{z}}_j\|_2 \cos(\theta_j) \in (-1, 1)$. Thus, all optimal $\widehat{\boldsymbol{\beta}}$'s satisfy $\widehat{\boldsymbol{\beta}}_j = \mathbf{0}$ for $j \in S^c$. □

B. Proofs of Theorems and Corollaries in Section IV

Proof of Theorem 4: Suppose the PDW construction succeeds. The proof of part (a) is given in Lemma 10. Further, in view of Lemma 10, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\text{PDW}}$ is an optimal solution of (10). Thus, all the optimal input vectors are supported on the set S, i.e., $\widehat{S} \subset S$, where $\widehat{S} = \{j : \widehat{\mathbf{u}}_j \neq 0\}$; thus, part (b) holds.

We show that the PDW construction succeeds with probability at least $1 - 2\exp(-T\delta^2/2)$ by showing that $\|\widehat{\mathbf{z}}_j\|_2 \leq 1$, for all $j \in S^c$. Here, $\widehat{\mathbf{z}}_j$ is determined in the step (c) of PDW construction. Let $\widehat{\boldsymbol{\beta}}_S$ be as in (36). By substituting $\mathbf{y}$ (given in (11)) and $\widehat{\mathbf{u}}_{S^c} = \mathbf{0}$ in (37), we obtain

$$\frac{1}{T}\begin{bmatrix}\mathbf{\Psi}_S^\mathsf{T}\mathbf{\Psi}_S & \mathbf{\Psi}_S^\mathsf{T}\widetilde{\mathbf{J}}_{S^c} \\ \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\mathbf{\Psi}_S & \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\mathbf{J}_{S^c}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S \\ \mathbf{0}\end{bmatrix} + \frac{1}{T}\begin{bmatrix}\mathbf{\Psi}_S^\mathsf{T} \\ \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\end{bmatrix}\mathbf{v} = \lambda_T\left[\begin{array}{c}\mathbf{0} \\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S \\ \hline \widehat{\mathbf{z}}_{S^c}\end{array}\right]. \qquad (41)$$

Using the second block equation of (41), solve for $\widehat{\mathbf{z}}_{S^c}$ as

$$\widehat{\mathbf{z}}_{S^c} = \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\mathbf{\Psi}_S\left[\frac{\mathbf{\Psi}_S^\dagger\mathbf{\Psi}_S}{\lambda_T T}(\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S)\right] + \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\left(\frac{\mathbf{v}}{\lambda_T T}\right), \qquad (42)$$

[5] The PDW construction is not an algorithm for solving (10): This is because to solve the problem in step (b) of PDW, we need to know S. However, PDW construction helps to prove theoretical results for the LASSO type problems.

where we used the fact $\boldsymbol{\Psi}_S = \boldsymbol{\Psi}_S \boldsymbol{\Psi}_S^\dagger \boldsymbol{\Psi}_S$. On the other hand, from the top block equation in (41), we have

$$\frac{1}{T}\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S(\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S) + \frac{1}{T}\boldsymbol{\Psi}_S^\mathsf{T}\mathbf{v} = \lambda_T \begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix}. \tag{43}$$

Pre-multiply both sides of the equality in (43) with $(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger$ and use the identity $\boldsymbol{\Psi}_S^\dagger = (\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger \boldsymbol{\Psi}_S^\mathsf{T}$ (see [36]) to get:

$$\boldsymbol{\Psi}_S^\dagger\boldsymbol{\Psi}_S(\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S) = -\boldsymbol{\Psi}_S^\dagger\mathbf{v} + T\lambda_T(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger \begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix}. \tag{44}$$

Let $\boldsymbol{\Gamma}_S = [\mathbf{I} - (\boldsymbol{\Psi}_S\boldsymbol{\Psi}_S^\dagger)]$. By substituting (44) in the first term of the second equality in (42), we can simplify $\widehat{\mathbf{z}}_{S^c}$ as

$$\widehat{\mathbf{z}}_{S^c} = \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}(\boldsymbol{\Psi}_S^\dagger)^\mathsf{T}\begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix} + \widetilde{\mathbf{J}}_{S^c}^\mathsf{T}\boldsymbol{\Gamma}_S\left(\frac{\mathbf{v}}{\lambda_T T}\right), \tag{45}$$

where we used the fact $(\boldsymbol{\Psi}_S^\dagger)^\mathsf{T} = \boldsymbol{\Psi}_S(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger$. Thus,

$$\widehat{\mathbf{z}}_j = \mathbf{J}_j^\mathsf{T}(\boldsymbol{\Psi}_S^\dagger)^\mathsf{T}\begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix} + \mathbf{J}_j^\mathsf{T}\boldsymbol{\Gamma}_S\left(\frac{\mathbf{v}}{\lambda_T T}\right), \quad \forall j \in S^c. \tag{46}$$

By the sub-multiplicative property of norms, for any $j \in S^c$,

$$\begin{aligned}\left\|\mathbf{J}_j^\mathsf{T}(\boldsymbol{\Psi}_S^\dagger)^\mathsf{T}\begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix}\right\|_2 &\leq \max_{j\in S^c}\|\mathbf{J}_j^\mathsf{T}(\boldsymbol{\Psi}_S^\dagger)^\mathsf{T}\|_2\left\|\begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix}\right\|_2\\ &\leq \frac{\alpha}{m^*}\|\mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\|_2 \leq \frac{\alpha}{m^*}\sum_{j\in S}\|\widehat{\mathbf{z}}_j\|_2 \leq \alpha.\end{aligned}$$

where $\alpha \leq 1$ is given in (23) and we used the fact that $\|\widehat{\mathbf{z}}_j\|_2 \leq 1$ (see Proposition 9), for $j \in S$, and $\|\mathbf{P}^\mathsf{T}\|_2 \leq 1$. As a result, from (46) and the preceding inequality, we have

$$\max_{j\in S^c}\|\widehat{\mathbf{z}}_j\|_2 \leq \alpha + \max_{j\in S^c}\left\|\mathbf{J}_j^\mathsf{T}\boldsymbol{\Gamma}_S\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2. \tag{47}$$

On the other hand, in light of Lemma 11, with the probability of at least $1-2\exp(-\delta^2T/2)$, we have $\|\mathbf{J}_j^\mathsf{T}\boldsymbol{\Gamma}_S\left(\mathbf{v}/\lambda_T T\right)\|_2 \leq 0.5(1-\alpha)$, for $\delta > 0$ and $j \in S^c$. Thus, $\max_{j\in S^c}\|\widehat{\mathbf{z}}_j\|_2 < 1$, establishing the strict dual feasibility condition.

Part (c): From Assumption (A2) and Proposition 1, we have

$$\boldsymbol{\Psi}_S^\dagger\boldsymbol{\Psi}_S = \operatorname{Blkdiag}(\mathbf{I}_n, \mathbf{I}_{t_S}, \boldsymbol{\Psi}_{S,[d-1:0]}^\dagger\boldsymbol{\Psi}_{S,[d-1:0]}), \tag{48}$$

where $t_S = (N-d+1)m^*$. Thus, we have

$$\mathbf{u}_{S,[0:N-d]}^* - \widehat{\mathbf{u}}_{S,[0:N-d]} = \boldsymbol{\Pi}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\boldsymbol{\Psi}_S(\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S), \tag{49}$$

where $\boldsymbol{\Pi}_{S,[0:N-d]} = [\mathbf{0}_{t_S\times n}\ \mathbf{I}_{t_S}\ \mathbf{0}_{t_S\times dm^*}]$ and

$$\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S = \begin{bmatrix}\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\\ \mathbf{u}_{S,[0:N-d]}^* - \widehat{\mathbf{u}}_{S,[0:N-d]}\\ \mathbf{u}_{S,[N-d+1:0]}^* - \widehat{\mathbf{u}}_{S,[N-d+1:0]}\end{bmatrix}. \tag{50}$$

From (49) and (44), it now follows that

$$\begin{aligned}\|\mathbf{u}_{S,[0:N-d]}^* - \widehat{\mathbf{u}}_{S,[0:N-d]}\|_\infty &\leq \|\boldsymbol{\Pi}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\mathbf{v}\|_\infty\\ &+ \lambda_T\left\|\boldsymbol{\Pi}_{S,[0:N-d]}(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S/T)^\dagger\right\|_\infty,\end{aligned} \tag{51}$$

where we used the fact $\|\widehat{\mathbf{z}}_S\|_\infty \leq 1$. The second term is deterministic. Instead, the first term is random, and, from Lemma 12, it is upper bounded by $\sigma/\sqrt{c_{\min}}(\sqrt{2\log(t_S)/T} + \delta)$ with probability at least $1-2\exp(-T\delta^2/2)$. Finally, the left-hand side of (51) equals $\max_{j\in S}\|\mathbf{u}_{j,[0:N-d]}^* - \widehat{\mathbf{u}}_{j,[0:N-d]}\|_\infty$. Putting the pieces together, we have the inequality in (25).

Part (d): By the triangle inequality, for all $j \in S$, we have

$$\begin{aligned}\|\mathbf{u}_{j:[0:N-d]}^*\|_\infty &= \|\mathbf{u}_{j:[0:N-d]}^* - \widehat{\mathbf{u}}_{j:[0:N-d]} + \widehat{\mathbf{u}}_{j:[0:N-d]}\|_\infty\\ &\leq \|\mathbf{u}_{j:[0:N-d]}^* - \widehat{\mathbf{u}}_{j:[0:N-d]}\|_\infty + \|\widehat{\mathbf{u}}_{j:[0:N-d]}\|_\infty\\ &\overset{(i)}{\leq} g_{\min}(\lambda_T, \boldsymbol{\Psi}) + \|\widehat{\mathbf{u}}_{j:[0:N-d]}\|_\infty,\end{aligned}$$

where (i) follows from part (c). Thus, $\|\widehat{\mathbf{u}}_{j:[0:N-d]}\|_\infty > 0$ if $\|\mathbf{u}_{j:[0:N-d]}^*\|_\infty > g_{\min}(\lambda_T, \boldsymbol{\Psi})$. This observation together with $\widehat{S} \subseteq S$ in part (a) implies that $\widehat{S} = S$.

Finally, the probability stated in the theorem is obtained by taking the union bound of the event where the dual feasibility holds and the event where ℓ_∞ bounds hold. □

Proof of Corollary 5: First, let us recall that $\widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]} = [\mathbf{I}_{n+t_S}\ \mathbf{0}_{(n+t_S)\times dm^*}]$. By proceeding similar to the steps outline in the proof of Theorem 4 (d), we get

$$\underbrace{\begin{bmatrix}\mathbf{x}_0^* - \widehat{\mathbf{x}}_0\\ \mathbf{u}_{S,[0:N-d]}^* - \widehat{\mathbf{u}}_{S,[0:N-d]}\end{bmatrix}}_{\boldsymbol{\beta}_{S,[0:N-d]}^* - \widehat{\boldsymbol{\beta}}_{S,[0:N-d]}} = \widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\boldsymbol{\Psi}_S(\boldsymbol{\beta}_S^* - \widehat{\boldsymbol{\beta}}_S).$$

Substituting this identity in (44) and followed by an application of triangle inequality yields us

$$\begin{aligned}\|\boldsymbol{\beta}_{S,[0:N-d]}^* - \widehat{\boldsymbol{\beta}}_{S,[0:N-d]}\|_2 &\leq \|\widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\mathbf{v}\|_2\\ &+ \lambda_T\left\|\widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S/T)^\dagger\begin{bmatrix}\mathbf{0}\\ \mathbf{P}^\mathsf{T}\widehat{\mathbf{z}}_S\end{bmatrix}\right\|_2.\end{aligned} \tag{52}$$

Using the facts that $\|\widehat{\mathbf{z}}_S\|_2 = \sqrt{\|\widehat{\mathbf{z}}_1\|_2^2 + \ldots + \|\widehat{\mathbf{z}}_{m^*}\|_2^2} \leq \sqrt{m^*}$, and $\mathbf{P}$ and $\widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}$ are permutation and selection matrices, respectively, the second term in (52) can be bounded above as $\lambda_T\|(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S/T)^\dagger\|_2\sqrt{m^*}$. By the hypothesis in the statement of the corollary, $\lambda_T\|(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S/T)^\dagger\|_2\sqrt{m^*} \leq \lambda_T\sqrt{m^*}/(c_{min})$.

We bound the first term in the upper bound in (52). Since $\mathbf{v} = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, we conclude that $\boldsymbol{\Sigma} = \sigma^2\widetilde{\boldsymbol{\Pi}}_S(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger\widetilde{\boldsymbol{\Pi}}_S^\mathsf{T}$ is the covariance of $\widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\mathbf{v}$. On the other hand, from Assumption (A2), $\|\boldsymbol{\Sigma}\|_2 \leq \sigma^2/(Tc_{min})$. Thus, by setting $t = 2(\sigma/\sqrt{c_{min}})(\sqrt{2c_1(n+t_S)/T} + \delta)$ in the first concentration result in Lemma 13, we have $\|\widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\mathbf{v}\|_2 \leq t$ with probability at least $1-\exp(-\delta^2T/2)$. The result in Corollary 5 follows by adding the upper bounds derived above. □

Proof of Theorem 6: From Theorem 4, $S = \widehat{S}$ holds with probability at least $1-4\exp(-T\delta^2/2)$. Thus, from (28), with the same probability, we have

$$\widehat{\boldsymbol{\beta}}_{\widehat{S},[0:N-d]}^{(OLS)} = \widehat{\boldsymbol{\beta}}_{S,[0:N-d]}^{(OLS)} = \widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}\boldsymbol{\Psi}_S^\dagger\mathbf{y}, \tag{53}$$

where $\widetilde{\boldsymbol{\Pi}}_S \triangleq \widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]} = [\mathbf{I}_{n+t_S}\ \mathbf{0}_{(n+t_S)\times dm^*}]$. Since $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\Psi}_S\boldsymbol{\beta}_S^*, \sigma^2\mathbf{I})$, from (53) and (19), we have

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_{\widehat{S},[0:N-d]}^{(OLS)}] = \widetilde{\boldsymbol{\Pi}}_{S,[0:N-d]}(\boldsymbol{\Psi}_S^\dagger\boldsymbol{\Psi}_S)\boldsymbol{\beta}_S^* = \boldsymbol{\beta}_{S,[0:N-d]}^*. \tag{54}$$

Let $\boldsymbol{\Sigma} \triangleq \sigma^2\widetilde{\boldsymbol{\Pi}}_S(\boldsymbol{\Psi}_S^\mathsf{T}\boldsymbol{\Psi}_S)^\dagger\widetilde{\boldsymbol{\Pi}}_S^\mathsf{T} \in \mathbb{R}^{n+t_S\times n+t_S}$. Then,

$$\widehat{\boldsymbol{\beta}}_{S,[0:N-d]}^{(OLS)} - \boldsymbol{\beta}_{S,[0:N-d]}^* \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{55}$$

We now upper bound $\|\mathbf{\Sigma}\|_2$. Recall from (18) and Proposition 1 (ii) that $\mathbf{\Psi}_S = [\mathbf{\Psi}_{S,[N:d]}\ \mathbf{\Psi}_{S,[d-1:0]}]$ and $\mathcal{R}(\mathbf{\Psi}_{S,[N:d]}) \cap \mathcal{R}(\mathbf{\Psi}_{S,[d-1:0]}) = \{0\}$. Then, from [27, Lemma D], we have

$$\widetilde{\mathbf{\Pi}}_S(\mathbf{\Psi}_S^{\mathsf{T}}\mathbf{\Psi}_S)^{\dagger}\widetilde{\mathbf{\Pi}}_S^{\mathsf{T}} = [(\mathbf{\Theta}\mathbf{\Psi}_{S,[N:d]})^{\mathsf{T}}(\mathbf{\Theta}\mathbf{\Psi}_{S,[N:d]})]^{\dagger}, \qquad (56)$$

where $\mathbf{\Theta} = \mathbf{\Theta}^2 = [\mathbf{I} - \mathbf{\Psi}_{S,[d-1:0]}\mathbf{\Psi}_{S,[d-1:0]}^{\dagger}]$. On the other hand, from Assumption (A2), we have $\|\mathbf{\Sigma}\|_2 \leq \sigma^2/(Tc_{\text{min}})$. Thu,s from the second concentration result in Lemma 13,

$$\left\|\boldsymbol{\beta}^*_{S,[0:N-d]} - \widehat{\boldsymbol{\beta}}^{(OLS)}_{S,[0:N-d]}\right\|_2 \leq \frac{4\sigma}{\sqrt{c_{min}}}\left\{\sqrt{\frac{(n+t_S)}{T}}\right\} + \frac{2\sigma}{\sqrt{c_{min}}}\left\{\sqrt{\frac{1}{T}\log\left(\frac{1}{\delta_1}\right)}\right\}, \qquad (57)$$

with probability at least $1-\delta_1$ for $\delta_1(0,1)$. The statement of the theorem follows by taking an union bound over the events where (57) and (53) hold. □

Proof of Theorem 7: Consider the auxiliary system $\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{b}_j u_j^*[k]$, where $j \in S^c$ and $u_j^*[k] = 0$, $k \geq N$. Let $\mathbf{x}[0] = 0$. Thus, $\mathbf{y} = \mathbf{J}_j\mathbf{u}_j^*$, where $\mathbf{J}_j$ is given by (7) and $\mathbf{y}$ and $\mathbf{u}_j^*$ as in (6). Let $\mathbf{\Psi}_S$ be as in (11), and consider

$$\begin{bmatrix}\widetilde{\mathbf{y}}[0]\\ \vdots \\ \widetilde{\mathbf{y}}[N]\end{bmatrix} \triangleq \mathbf{\Psi}_S^{\dagger}\begin{bmatrix}\mathbf{y}[0]\\ \vdots \\ \mathbf{y}[N]\end{bmatrix} = \mathbf{\Psi}_S^{\dagger}\mathbf{J}_j\mathbf{u}_j^* = \mathbf{\Psi}_S^{\dagger}\mathbf{J}_j\begin{bmatrix}u_j^*[0]\\ \vdots \\ u_j^*[N]\end{bmatrix}, \qquad (58)$$

By assumption we have $\text{nRank}\mathcal{Z}_S = n + m^*$. Thus, for all $z \notin \text{spec}(A)$, $\mathcal{G}_S[z]$ has full column rank and $\mathcal{G}_S^+[z] = [\mathcal{G}_S[z]^{\mathsf{T}}\mathcal{G}_S[z]]^{-1}\mathcal{G}_S[z]^{\mathsf{T}}$ and $\mathcal{G}_S^+[z]\mathcal{G}_S[z] = z^{-d}\mathbf{I}$; see [9, Theorem 1]. Let $\widetilde{\mathbf{y}}[z]$ be the $\mathcal{Z}$-transform of $\{\widetilde{\mathbf{y}}[k]\}_{k=0}^{\infty}$. Then by using the construction given in [37, pp. 49-50] and the uniqueness of pseudo inverse [27], we have $\widetilde{\mathbf{y}}[z] = z^{-d}\mathcal{H}[z]u_j^*[z]$, where $\mathcal{H}_j[z] = \mathcal{G}_S^+[z]\mathcal{G}_j[z]$, for all $z \notin \text{spec}(A)$.

On the one hand from Parsevel's theorem, we have

$$\begin{aligned}\sqrt{\sum_{k=0}^{\infty}\|\widetilde{\mathbf{y}}[k]\|_2^2} &= \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\|\widetilde{\mathbf{y}}[e^{j\omega}]\|_2^2 d\omega}\\ &= \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\|e^{-dj\omega}\mathcal{H}_j[e^{j\omega}]u_j^*[e^{j\omega}]\|_2^2 d\omega}\\ &\leq \sup_{\{\omega\in[-\pi,\pi]\}}\|\mathcal{H}_j[e^{j\omega}]\|_2\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}|u_j^*[e^{j\omega}]|_2^2 d\omega}\\ &= \sup_{\{z\in\mathbb{C}:|z|=1\}}\|\mathcal{H}_j[z]\|_2\|\mathbf{u}_j^*\|_2. \qquad (59)\end{aligned}$$

For the last inequality, we used Parsevel's theorem and $u^*[k] = 0$, for $k > N$. On the other hand, from (58) and (59), we have

$$\begin{aligned}\|\mathbf{\Psi}_S^{\dagger}\mathbf{J}_j\|_2 &= \sup_{\|\mathbf{u}_j^*\|_2=1}\|\mathbf{\Psi}_S^{\dagger}\mathbf{J}_j\mathbf{u}_j^*\|_2 = \sup_{\|\mathbf{u}_j^*\|_2=1}\sqrt{\sum_{k=0}^{N}\|\widetilde{\mathbf{y}}[k]\|_2^2}\\ &\leq \sup_{\|\mathbf{u}_j^*\|_2=1}\sqrt{\sum_{k=0}^{\infty}\|\widetilde{\mathbf{y}}[k]\|_2^2} \leq \sup_{\{z\in\mathbb{C}:|z|=1\}}\|\mathcal{H}_j[z]\|_2.\end{aligned}$$

Thus $\max_{\{j\in S^c\}}\sup_{\{z\in\mathbb{C}:|z|=1\}}\|\mathcal{H}[z]\|_2 \leq \alpha/m^*$ implies that $\|\mathbf{\Psi}_S^{\dagger}\mathbf{J}_j\|_2 \leq \alpha/m^*$. The proof is now complete. □

C. Auxiliary lemmata

***Lemma* 11.** *With the notation and assumptions in Theorem 4, we have* $\mathbb{P}[\max_{j\in S^c}\left\|\mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\left(\mathbf{v}/\lambda_T T\right)\right\|_2 \geq 0.5(1-\alpha)] \leq 2\exp(-T\delta^2/2)$*, where* $\mathbf{\Gamma}_S = [\mathbf{I} - \mathbf{\Psi}_S\mathbf{\Psi}_S^{\dagger}]$ *and* $\delta > 0$.

Proof. Let $\widetilde{\alpha} = 0.5(1-\alpha)$, with $\alpha \in [0,1)$. Consider

$$\begin{aligned}&\mathbb{P}\left[\max_{j\in S^c}\left\|\mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq \widetilde{\alpha}\right]\\ &\qquad\leq \sum_{j\in S^c}\mathbb{P}\left[\left\|\mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq \widetilde{\alpha}\right]. \qquad (60)\end{aligned}$$

Because $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, it follows that $\mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\left(\mathbf{v}/\lambda_T T\right) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_j)$ with $\mathbf{\Sigma}_j = \mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\mathbf{\Gamma}_S^{\mathsf{T}}\mathbf{J}_j/(\lambda_T^2T^2)$. Furthermore,

$$\|\mathbf{\Sigma}_j\|_2 = \frac{1}{\lambda_T^2T^2}\|\mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\|_2^2 \leq \frac{1}{\lambda_T^2T^2}\|\mathbf{J}_j^{\mathsf{T}}\|_2^2 \leq \frac{C^2T}{\lambda_T^2T^2}. \qquad (61)$$

The first inequality follows because $\mathbf{\Gamma}_S$ is a projection matrix and for the last inequality from the normalization assumption (A1). Invoking Lemma 13, we bound the inequality in (60) as

$$\sum_{j\in S^c}\mathbb{P}\left[\left\|\mathbf{J}_j^{\mathsf{T}}\mathbf{\Gamma}_S\left(\frac{\mathbf{v}}{\lambda_T T}\right)\right\|_2 \geq \widetilde{\alpha}\right] \leq \sum_{j\in S^c}c_N\exp\left(-\frac{\widetilde{\alpha}^2\lambda_T^2T}{8\sigma^2C^2}\right),$$

where $c_N = 5^{N+1}$. The right side term can be simplified as

$$\exp\left((N+1)\log(5) + \log(m-m^*) - \frac{\widetilde{\alpha}^2\lambda_T^2T}{8\sigma^2C^2}\right) \qquad (62)$$

Substituting λ_T (see (24)) and $\widetilde{\alpha} = 0.5(1-\alpha)$ in (62), and simplifying it gives us the required bound. □

***Lemma* 12.** *With the notation and assumptions stated in Theorem 4, for* $\delta \in [0,1)$*, we have* $\mathbb{P}[\|\mathbf{\Pi}_{S,[0:N-d]}\mathbf{\Psi}_S^{\dagger}\mathbf{v}\|_\infty \geq \sigma/\sqrt{c_{\text{min}}}(\sqrt{2\log(t_S)/T} + \delta)] \leq 2\exp(-T\delta^2/2)$.

Proof. Let $\mathbf{e}_l$ is the l^{th} standard basis vector in $\mathbb{R}^{t_S}$, and $t_S = (N-d+1)m^*$. Define $a_l = \mathbf{e}_l^{\mathsf{T}}\mathbf{\Pi}_{S,[0:N-d]}\mathbf{\Psi}_S^{\dagger}\mathbf{v}$ to be the l^{th} entry of $\mathbf{\Pi}_{S,[0:N-d]}\mathbf{\Psi}_S^{\dagger}\mathbf{v}$. Then, because $\mathbf{\Pi}_{S,[0:N-d]} = [\mathbf{0}_{t_S\times n}\ \mathbf{I}_{t_S}\ \mathbf{0}_{t_S\times dm^*}]$, it follows that $\|\mathbf{\Pi}_{S,[0:N-d]}\mathbf{\Psi}_S^{\dagger}\mathbf{v}\|_\infty = \max_{l\in 1\ldots t_S}|a_l|$. Thus, for any $\kappa \geq 0$, we have the bound:

$$\mathbb{P}[\max_{\{l\in 1\ldots t_S\}}|a_l| \geq \kappa] \leq \sum_{l=1}^{t_S}\mathbb{P}[|a_l| \geq \kappa]. \qquad (63)$$

We bound terms on the right-hand side by invoking standard concentration results. For compactness, let $\mathbf{\Pi}_{S,} = \mathbf{\Pi}_{S,[0:N-d]}$. Because $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, we have $a_l \sim \mathcal{N}(\mathbf{0}, \sigma_l^2)$, where

$$\begin{aligned}\sigma_l^2 = \sigma^2\mathbf{e}_l^{\mathsf{T}}\mathbf{\Pi}_S\mathbf{\Psi}_S^{\dagger}(\mathbf{\Pi}_S\mathbf{\Psi}_S^{\dagger})^{\mathsf{T}}\mathbf{e}_l &\leq \sigma^2\lambda_{\text{max}}(\mathbf{\Pi}_S\mathbf{\Psi}_S^{\dagger}(\mathbf{\Pi}_S\mathbf{\Psi}_S^{\dagger})^{\mathsf{T}})\\ &= \sigma^2\|\mathbf{\Pi}_S(\mathbf{\Psi}_S^{\mathsf{T}}\mathbf{\Psi}_S)^{\dagger}\mathbf{\Pi}_S^{\mathsf{T}}\|_2\\ &\leq \sigma^2\|\widetilde{\mathbf{\Pi}}_S(\mathbf{\Psi}_S^{\mathsf{T}}\mathbf{\Psi}_S)^{\dagger}\widetilde{\mathbf{\Pi}}_S^{\mathsf{T}}\|_2\\ &\leq \sigma^2/(Tc_{\text{min}}). \qquad (64)\end{aligned}$$

where $\widetilde{\mathbf{\Pi}}_S = [\mathbf{I}_{n+t_S}\ \mathbf{0}_{(n+t_S)\times dm^*}]$. The second inequality follows from interlacing property of singular values. The final inequality is showed in the proof of Theorem 6.

Since a_l is Gaussian, from [30, page 22] and (64), we have $\mathbb{P}[|z_l| \geq \kappa] \leq \exp(-\kappa^2/(2\sigma_l^2)) \leq \exp(-\kappa^2Tc_{\text{min}}/(2\sigma^2))$. Substituting this inequality in (65), we find that

$$\mathbb{P}[\max_{l\in 1\ldots t_S}|z_l| \geq \kappa] \leq \exp\left(\log(t_S) - \frac{\kappa^2Tc_{\text{min}}}{2\sigma^2}\right). \qquad (65)$$

The result follows by letting $\kappa = \sigma/\sqrt{c_{\min}}(\sqrt{2\log(t_S)/T}+\delta)$ and simplifying terms in the exponential term. □

***Lemma* 13.** *Let $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} \in \mathbb{R}^{l\times l}$ is a positive definite matrix. Then, $\mathbb{P}[\|\mathbf{p}\|_2 \geq t] \leq 5^l \exp(-t^2/(8\|\mathbf{\Sigma}\|_2))$. Furthermore, $\|\mathbf{p}\|_2 \leq 4\sqrt{\|\mathbf{\Sigma}\|_2 l} + 2\sqrt{\|\mathbf{\Sigma}\|_2 \log(1/\delta)}$ with probability at least $1-\delta$ for $\delta \in (0,1)$.*

Proof. Follows from Lemma 8.2 and Theorem 8.3 in [38]. □

References

[1] M. Ghorbaniparvar. Survey on forced oscillations in power system. *Journal of Modern Power Systems & Clean Energy*, 5(5):671–682, 2017.

[2] B. Wang and K. Sun. Location methods of oscillation sources in power systems: a survey. *Journal of Modern Power Systems & Clean Energy*, 5(2):151–159, 2017.

[3] T. Huang. et al. A synchrophasor data-driven method for forced oscillation localization under resonance conditions. *IEEE Transactions on Power Systems*, 35(5):3927–3939, 2020.

[4] S. C. Chevalier, V. Petr, and K. Turitsyn. Using effective generator impedance for forced oscillation source location. *IEEE Transactions on Power Systems*, 33(6):6264–6277, 2018.

[5] S. Maslennikov, B. Wang, Q. Zhang, F. Ma, X. Luo, K. Sun, and E. Litvinov. A test cases library for methods locating the sources of sustained oscillations. In *2016 IEEE Power and Energy Society General Meeting*, pages 1–5. 2016.

[6] K. Lounici, M. Pontil, S. Van De Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164 – 2204, 2011.

[7] N. Simon and R. Tibshirani. Standardization and the group LASSO penalty. *Statistica Sinica*, 22(3):983–1001, 2012.

[8] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group LASSO for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[9] S. Kirtikar, H. Palanthandalam-Madapusi, E. Zattoni, and D. S. Bernstein. *l*-delay input and initial-state reconstruction for discrete-time linear systems. *Circuits Syst Signal Process*, 30:233–262, 2011.

[10] S. Boyd et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[11] S. Chevalier, P. Vorobev, and K. Turitsyn. A Bayesian approach to forced oscillation source location given uncertain generator parameters. *IEEE Transactions on Power Systems*, 34(2):1641–1649, 2019.

[12] U. Agrawal, J. W. Pierre, J. Follum, D. Duan, D. Trudnowski, and M. Donnelly. Locating the source of forced oscillations using PMU measurements and system model information. In *2017 IEEE Power Energy Society General Meeting*, pages 1–5, 2017.

[13] N. Zhou, M. Ghorbaniparvar, and S. Akhlaghi. Locating sources of forced oscillations using transfer functions. In *2017 IEEE Power and Energy Conference at Illinois (PECI)*, pages 1–8, 2017.

[14] Y. Meng, Z. Yu, N. Lu, and D. Shi. Time series classification for locating forced oscillation sources. *IEEE Transactions on Smart Grid*, 12(2):1712–1721, 2021.

[15] F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, 2013.

[16] F. Dörfler, F. Pasqualetti, and F. Bullo. Continuous-time distributed observers with discrete communication. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):296–304, 2013.

[17] M. Luan, D. Gan, Z. Wang, and H. Xin. Application of unknown input observers to locate forced oscillation source. *International Transactions on Electrical Energy Systems*, 29(9), 2019.

[18] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, 2014.

[19] M. B. Wakin, B. M. Sanandaji, and T. L. Vincent. On the observability of linear systems from random, compressive measurements. In *49th IEEE Conference on Decision and Control*, pages 4447–4454, 2010.

[20] G. Joseph and C. R. Murthy. Measurement bounds for observability of linear dynamical systems under sparsity constraints. *IEEE Transactions on Signal Processing*, 67(8):1992–2006, 2019.

[21] S. Sefati, N. J. Cowan, and R. Vidal. Linear systems with sparse inputs: Observability and input recovery. In *2015 American Control Conference (ACC)*, pages 5251–5257, 2015.

[22] S. M. Fosson, F. Garin, S. Gracy, A. Y. Kibangou, and D. Swart. Input and state estimation exploiting input sparsity. In *2019 18th European Control Conference (ECC)*, pages 2344–2349, 2019.

[23] S. Z. Yong, M. Zhu, and E. Frazzoli. A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems. *Automatica*, 63:321–329, 2016.

[24] S. Sundaram and C.N. Hadjicostis. Distributed function calculation via linear iterative strategies in the presence of malicious agents. *IEEE Transactions on Automatic Control*, 56(7):1495–1508, 2011.

[25] Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.

[26] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[27] A. Ansari and D. S. Bernstein. Deadbeat unknown-input state estimation and input reconstruction for linear discrete-time systems. *Automatica*, 103:11–19, 2019.

[28] B. D. O. Anderson and M. Deistler. Properties of zero-free spectral matrices. *IEEE Trans. on Automatic Control*, 54(10):2365–2375, 2009.

[29] H. Liu and J. Zhang. On the ℓ_1-ℓ_p regularized regression. Technical report, Department of Statistics, Carnegie Mellon University, 2009.

[30] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[31] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

[32] R. Anguluri, L. Sankar, and O. Kosut. Localization and estimation of unknown forced inputs: A group LASSO approach. *arXiv preprint arXiv:2201.07907*, 2022.

[33] J. H. Chow and K. W. Cheung. A toolbox for power system dynamics and control engineering education and research. *IEEE Transactions on Power Systems*, 7(4):1559–1564, 1992.

[34] A. K. Singh et al. Report on the 68-bus, 16-machine system. *IEEE PES Task Force on Benchmark System for Stability Controls. Ver*, 3, 2013.

[35] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[36] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. New York: Springer-Verlag, 2003.

[37] A. Ansari. *Input and State Estimation for Discrete-Time Linear Systems with Application to Target Tracking and Fault Detection*. PhD dissertation, The University of Michigan, 2018.

[38] A. Rinaldo. Advanced statistical theory, lecture 8, 2019. "URL: http://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures/Feb21_Shenghao.pdf.

Rajasekhar Anguluri (Member, IEEE) received the B.Tech. degree in electrical engineering from the National Institute of Technology Warangal, India, in 2013, and both the M.S. degree in statistics and the Ph.D. degree in mechanical engineering from the University of California at Riverside, CA, USA, in 2019. He is currently a postdoctoral research scholar with the School of Electrical, Computer, and Energy Engineering at Arizona State University, Tempe, AZ, USA. His current research interests include high-dimensional statistics, statistical signal processing, stochastic control, and power systems.

Lalitha Sankar (Senior Member, IEEE) received the bachelor's degree from the Indian Institute of Technology Bombay, Mumbai, India, the master's degree from the University of Maryland, College Park, MD, USA, and the Doctorate degree from Rutgers University, New Brunswick, NJ, USA. She is an Associate Professor with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA. She is currently with an NSF-Funded Institute focused on using data science to enable real-time integration of synchrophasor data into grid operations. She was the recipient of the NSF CAREER Award.

Oliver Kosut (Member, IEEE) received B.S. degree in electrical engineering and mathematics from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2004 and the Ph.D. degree in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2010. Since 2012, he has been a Faculty Member with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA, where he is an Associate Professor. From 2010 to 2012, he was a Postdoctoral Research Associate with Laboratory for Information and Decision Systems, MIT. His research interests include information theory, cybersecurity, and power systems. He was the recipient of the NSF CAREER Award in 2015.