2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Multi-Task System Identification of Similar Linear Time-Invariant Dynamical Systems

Yiting Chen, Ana M. Ospina, Fabio Pasqualetti, and Emiliano Dall'Anese

Abstract— **This paper presents a system identification framework – inspired by multi-task learning – to estimate the dynamics of a given number of linear time-invariant (LTI) systems jointly by leveraging structural similarities across the systems. In particular, we consider LTI systems that model networked systems with similar connectivity, or LTI systems with small differences in their matrices. The system identification task involves the minimization of the least-squares (LS) fit for individual systems, augmented with a regularization function that enforces structural similarities. The proposed method is particularly suitable for cases when the recorded trajectories for one or more LTI systems are not sufficiently rich, leading to ill-conditioning of LS methods. We analyze the performance of the proposed method when the matrices of the LTI systems feature a common sparsity pattern (i.e., similar connectivity), and provide simulations based on real data for the estimation of the brain dynamics. We show that the proposed method requires a significantly smaller number of fMRI scans to achieve similar error levels of the LS.**

I. Introduction

System identification is a core task where the model of dynamical systems is estimated based on observed inputs and states [1], [2]. In particular, identification of linear time-invariant (LTI) systems is a well-investigated problem that has recently received renewed attention due to lines of research in the context of data-driven control and optimization (see, for example, the representative works in [3]–[8]).

When the observation of the state is noise-free, the LTI system matrices can be estimated by leveraging the Willems' Fundamental Lemma, provided that the recorded trajectory satisfies the persistency of excitation (PE) condition as discussed in, e.g., [9], [10]. On the other hand, when process noise or disturbances enter the LTI system, several existing works focus on the asymptotic and finite time estimation errors and sample complexity of the least squares (LS) estimator; see, for example, the representative works [11]–[16] and pertinent references therein. Additionally, regularized system identification methods are investigated in, e.g., [2], [17], [18]; a low-order linear system identification via regularized regression is considered in [19]. These regularized identification methods allow one to add a prior on the system matrices, and to strike a balance between LS fit and model complexity [20], [21].

The performance of the LS estimator hinges on the availability of a recorded trajectory that is sufficiently rich to render the LS well conditioned. In this paper, we are interested in cases where the LS method is ill-conditioned. In particular, we consider the task of estimating the system matrices of $N > 1$ LTI systems, in cases where we do not have sufficiently long (and sufficiently rich) recorded trajectories for at least one of the systems (or for some of the systems). Accordingly, the question posed in this paper is as follows: *is it possible to leverage "similarities" among the N systems to obtain accurate estimates of the system matrices, even if the LS is ill-conditioned? In particular, if one has only a few measurements for the i-th system, can one use recorded data from the other LTI systems to improve the estimation error?*

In this direction, [22] considered estimating the matrices of a linear system from samples generated by a "similar" one; in [22], an LTI system is considered "similar" if its matrices are perturbed versions of a given matrix. Recently, [23] considered a setup where the matrix norm of the difference between the matrices of LTI systems is small. In this paper, we expand the notion of "structural similarity" to account for additional properties that the N systems may have in common, and propose a new system identification approach that leverages and cross-fertilizes core tools investigated in the context of multi-task learning [24]–[26], statistical learning [20], and regularized identification methods [2], [27].

We first consider the case where the N LTI systems model networked systems with a *similar connectivity (sub-)graph*; this implies that the matrices of the LTI systems feature a *common sparsity* (i.e., the system matrices have zeros in a common set of entries). With this model, we formulate the multi-task system identification task as a *regularized LS* problem where we minimize the LS fit for each of the LTI systems plus a regularization function that enforces a common sparsity pattern [28], [29]. By appropriately tuning (typically via cross-validation [21]) the weight assigned to the regularization function, one can find a balance between fitting of the recorded data and model complexity. We analyze the estimation error of the proposed multi-task system identification approach, with respect to the true matrices of the LTI systems and with respect to an "oracle;" the latter represents the best achievable estimation when considering a (group) sparse model, under a given model compatibility condition [30].

Next, we explain how the proposed multi-task system identification can be adjusted to account for additional struc-

Y. Chen, A. M. Ospina, and E. Dall'Anese are with the Department of Electrical, Computer and Energy Engineering, University of Colorado Boulder; emails: {yiting.chen-1, ana.ospina, emiliano.dallanese}@colorado.edu. F. Pasqualetti is with the Department of Mechanical Engineering, University of California Riverside; email: fabiopas@engr.ucr.edu.

This work was supported in part by the National Science Foundation (NSF) through the awards 1941896 and 1926829, and by the Army Research Office (ARO) through the award W911NF1910360.

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

tural similarities. In particular, we provide approaches to deal with cases where the matrix *feature a small heterogeneity* (i.e., the matrix difference is small, as in [23]), and where some of the *matrices of the LTI systems can be expressed as linear combinations* of each other. In this case, we resort to regularization functions that penalize large matrix deviations and functions that are inspired by nuclear norm minimization [31]–[33].

We demonstrate the effectiveness of the proposed multi-task system identification method using: (i) synthetic LTI systems that feature structural similarities, and (ii) real data from the Human Connectome Project (HCP), where blood-oxygen-level-dependent (BOLD) signals are obtained from resting state functional magnetic resonance imaging (fMRI) [34], [35]. For the latter, we show that the proposed method requires a significantly smaller number of fMRI scans to achieve the same error of the LS by simply assuming that the underlying functional or structural connectivity of brain parcellations is similar across subjects. We also consider the case where only a few fMRI readings are available for one subject, showing the ability to "transfer information" from the dynamics of the other subjects.

II. Preliminaries and System Identification Setup

Consider N linear time-invariant (LTI) systems[1]

$$x_i(t+1) = A_i x_i(t) + B_i u_i(t) + w_i(t), \quad x_i(0) \in \mathbb{R}^n, \tag{1}$$

with $i \in [N]$ the system index and $t \in \mathbb{N}$ the time index, $A_i \in \mathbb{R}^{n\times n}$ and $B_i \in \mathbb{R}^{n\times p}$, and where $x_i(t) \in \mathbb{R}^n$, $u_i(t) \in \mathbb{R}^p$, and $w_i(t) \in \mathbb{R}^n$ are the state, input and process noise, respectively, of the ith system. Assume that, for each system, the input $u_i(t)$ and state $x_i(t)$ can be measured, and $B_i \in \mathbb{R}^{n\times p}$ is known; on the other hand, the system matrix is unknown and the disturbance $w_i(t)$ cannot be measured[2].

For the i-th system, suppose that one has access to one trajectory $\{x_i(\tau), u_i(\tau)\}_{\tau=1}^{P_i+1}$, for some $P_i \in \mathbb{N}$, for the state and the inputs. With these measurements, the system matrices can be estimated using the following LS criterion:

$$\min_{A_i \in \mathbb{R}^{n\times n}} \mathcal{L}_i(A_i), \tag{2}$$

where $\mathcal{L}_i(A_i) := \sum_{\tau=1}^{P_i} \|x_i(\tau+1) - A_i x_i(\tau) - B_i u_i(\tau)\|_2^2$; the LS problem (2) is solved for each of the N systems independently. The LS estimator (2) has been extensively studied in the literature, especially when the recorded data render the LS (2) well conditioned [11]–[14], [19].

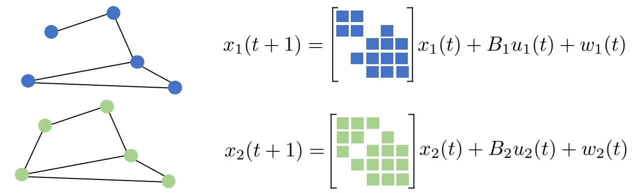

$x_1(t+1) =$
$x_1(t) + B_1 u_1(t) + w_1(t)$
$x_2(t+1) =$
$x_2(t) + B_2 u_2(t) + w_2(t)$

Fig. 1. Example of network systems with similar connectivity. the two matrices A_1 and A_2 feature zeros on a common set of entries.

In this paper, we are interested in cases where the LS problem (2) is ill-conditioned for some of the N LTI systems; this may be due to recorded trajectories that are not sufficiently rich, or simply not long enough. In this case, the question we pose in this paper pertains to whether it is possible to leverage "similarities" among the N systems to obtain accurate estimates of the system matrices, even if the LS is ill-conditioned for some systems. We consider the following structural similarities across the LTI systems:

(s1) The matrices $A_1, \ldots A_N$ have zeros in the same entries; i.e., $(A_1)_{ij} = (A_2)_{ij} = \ldots = (A_N)_{ij} = 0$ for some entries (i, j).

(s2) For any pair A_i, A_j, $i, j \in [N]$, there exists $\epsilon > 0$ such that $\|A_i - A_j\|_F^2 \leq \epsilon$.

(s3) For the subset of systems $i \in \mathcal{C}, \mathcal{C} \subseteq [N]$, there exists $\{\alpha_{i,j} \in \mathbb{R}\}$ such that $A_i = \sum_{j=1, j\neq i}^{N} \alpha_{ij} A_j$.

We note that (s1) naturally models LTI systems that describe the dynamics of networked systems with similar connectivity; as a concrete example, (s1) emerges from a similar functional or structural connectivity of the brain network across different individuals [34], [35]. Similarity, (s2) models the case where the norm of the matrix difference $A_i - A_j$ is small; this is in line with the models considered in [22], [23]. Finally, (s3) models the case where the matrix A_i of the i-th system can be expressed as a linear combination of some of the other matrices $\{A_j\}_{j=1, j\neq i}^{N}$; as an example, this model may be applicable to traffic flows and mobility-on-demand services (see, e.g., [36]), where the LTI systems (1) model the evolution of the density of vehicles in given geographical areas over given periods of the day.

Given the models (s1)-(s3), we consider estimating the matrices $\{A_i\}_{i\in[N]}$ jointly by solving the following system identification problem:

$$\min_{\{A_k\}_{k\in[N]}} \sum_{i=1}^{N} \mathcal{L}_i(A_i) + \lambda \mathcal{R}(A_1, \ldots, A_N), \tag{3}$$

where we recall that $\mathcal{L}_i(A_i)$ is the LS fit for the ith system and, in the spirit of regularized regression methods [2], [20], [27], $(A_1, \ldots, A_N) \mapsto \mathcal{R}(A_1, \ldots, A_N)$ is a lower-semicontinuous convex function that promotes the prior specified by (s1)–(s3), and $\lambda > 0$ is a tuning parameter. Problem (3) is inspired by multi-task learning methods [24]–[26], [37], where learning tasks are performed simultaneously (in

[1] **Notation**: We denote by $\mathbb{N}$ and $\mathbb{R}$ the set of natural numbers and the set of real numbers, respectively, and define $[n] = \{1, 2, \ldots, n\}$. Given $S \subset [n]$, $|S|$ is the cardinality of S. We let $^\top$ denote transposition. For a given column vector $x \in \mathbb{R}^n$, $\|x\|_2$ is the Euclidean norm and $\|x\|_1$ denotes the ℓ_1 norm; for a matrix $X \in \mathbb{R}^{n\times m}$, $\|X\|_F$ denotes the Frobenious norm and $\|X\|_*$ the nuclear norm. Moreover, $(x)_i$ refers to the entry i of the vector x, $(X)_{ij}$ to the entry (i, j) of the matrix X, and $\text{vec}(X)$ is a $mn \times 1$ vector stacking the columns of X. Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $\nabla f(x)$ denotes the gradient of f at x (taken to be a column vector). Given a closed convex set $C \subseteq \mathbb{R}^n$, $\text{proj}_C : \mathbb{R}^n \to \mathbb{R}^n$ denotes the Euclidean projection of y onto C, namely $\text{proj}_C(y) := \arg\min_{v\in C} \|y - v\|$. Given a lower-semicontinuous convex function $g : \mathbb{R}^n \to \mathbb{R}$, the proximal operator is defined as $\text{prox}_{\lambda g}(y) := \arg\min_{x\in\mathbb{R}^n} g(x) + \frac{1}{2\lambda}\|x - y\|_2^2$.

[2] We note that, while we focus on the estimation of the matrix A_i for notation simplicity and to streamline exposition, the techniques presented in the paper are straightforwardly applicable to the case where both A_i and B_i are estimated from data.

our case, the LS fitting) while exploiting commonalities on the parameters that are learned [26, Definition 1]. In the ensuing sections, we will explain specific choices for the regularization function $\mathcal{R}(A_1, \dots, A_N)$; we start with the case where the N LTI systems feature a common connectivity.

III. Systems with similar connectivity

A. Multi-task System Identification

In this section, we consider the case where the system matrices $A_1, \dots A_N$ have zeros in a common set of entries; i.e., $(A_1)_{ij} = (A_2)_{ij} = \dots = (A_N)_{ij} = 0$ for a subset of indices $i \in [n]$ and $j \in [n]$. In the statistical learning literature, this structural similarity leads to a setup where the unknowns $\{A_1, \dots, A_N\}$ feature a *group sparsity* [28], [29]. Leveraging the technical approach of [28], [29], we then formulate the multi-task system identification problem for LTI systems with similar connectivity as follows:

$$\min_{\{A_k\}_{k\in[N]}} \sum_{i=1}^{N} \sum_{\tau=1}^{P_i} \|x_i(\tau+1) - A_i x_i(\tau) - B_i u_i(\tau)\|_2^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} \|[(A_1)_{ij}, (A_2)_{ij}, \dots, (A_N)_{ij}]\|_2, \quad (4)$$

where $\lambda \geq 0$ and the function $\mathcal{R}(A_1, \dots, A_N) = \sum_{i=1}^{N} \sum_{j=1}^{N} \|[(A_1)_{ij}, (A_2)_{ij}, \dots, (A_N)_{ij}]\|_2$ is utilized to enforce group sparsity; that is, the solution of (4) is such that either the whole vector $[(A_1)_{ij}, (A_2)_{ij}, \dots, (A_N)_{ij}]$ is zero or not. The parameter λ in (3) strikes a balance between the LS fit (in our case, the LS fit for individual LTI systems) and a number of vectors $[(A_1)_{ij}, (A_2)_{ij}, \dots, (A_N)_{ij}]$ that are set to zero (as shown shortly).

Problem (4) is an unconstrained convex program; given the composite cost, we consider a proximal-gradient method (with line search) for solving (4) (see, e.g., [38], [39]). The proximal-gradient method is tabulated as Algorithm 1.

Algorithm 1 Proximal gradient method for systems with similar connectivity

```
Given: $\hat{A}_1^{(0)}, \cdots, \hat{A}_N^{(0)}, \eta^{(0)}$, $\beta \in (0,1)$, and $\lambda$.
Repeat: $m = 0, 1, 2, \dots$ until convergence
  [S1] $\alpha \leftarrow \eta^{(m)}$.
  [S2] Proximal-gradient with line search:
    [S2.1] $Z_i = \hat{A}_i^{(m)} - \alpha \nabla \mathcal{L}_i(\hat{A}_i^{(m)})$, $i \in [N]$
    [S2.2] Let $z_{jk} := [(Z_1)_{jk}, (Z_2)_{jk}, \dots, (Z_N)_{jk}]$.
           For all $j, k \in [n]$, compute:
             $y_{jk} = \frac{z_{jk}}{\|z_{jk}\|_2} \max(\|z_{jk}\|_2 - \alpha\lambda, 0)$.   (5)
    [S2.3] Form matrices $\{Y_\ell\}_{\ell \in [N]}$ as $(Y_\ell)_{jk} = (y_{jk})_\ell$.
    [S2.4] Break if: $\sum_{i=1}^{N} \mathcal{L}_i(Y_i) \leq \frac{1}{2\lambda}\|Y_i - \hat{A}_i^{(m)}\|_F^2$
           $+ \sum_{i=1}^{N} \left(\mathcal{L}_i(\hat{A}_i^{(m)}) + \nabla \mathcal{L}_i(\hat{A}_i^{(m)})^\top (Y_i - \hat{A}_i^{(m)})\right)$
    [S2.5] Update $\alpha \leftarrow \beta\alpha$.
  [S3] $\eta^{(m+1)} \leftarrow \alpha, \hat{A}_i^{(m+1)} \leftarrow Y_i$, $i \in [N]$.
```

We note that the step [S2.3] is in fact a closed-form expression for the proximal map:

$$\{Y_i\}_{i\in[N]} = \text{prox}_{\alpha\lambda\mathcal{R}}(\{Z_i\}_{i\in[N]}), \quad (6)$$

and it involves n^2 parallel computations as in (5). The convergence behavior of Algorithm 1 can be readily analyzed by leveraging the results in [38, Chapter 2]. Moreover, Algorithm 1 can be converted into a "standard" proximal-gradient method if the line search is not performed [40].

From the thresholding operation (5), it is clear that increasing λ has the effect of forcing a higher number of entries of the system matrices to be zero. Unfortunately, tuning λ is not an easy task, and cross-validation procedures are typically utilized to find the value of λ such that the estimated matrices yield the lowest error on test data; see, for example [20], [21].

B. Analysis

In this section, we analyze the performance of the multi-task system identification method. The performance of regression problems with sparsity-enforcing regularization terms is oftentimes compared against an "oracle" that represents the *best achievable estimation* when considering a (group) sparse model, under a given model compatibility condition [30]. We will provide error bounds with respect to both the oracle and the true system matrices.

To simplify the notation, we outline the results for the case where $P_i = P$ for all $i \in [N]$ (though, similar results hold when the $\{P_i\}$ are different). Let $A^S \in \mathbb{R}^{n\times n}$ be a matrix formed by selecting the entries of $A \in \mathbb{R}^{n\times n}$ indexed by $S \subset [n] \times [n]$ and setting zero to the other entries. For N matrices $A_i \in \mathbb{R}^{n\times n}$, $i \in [N]$, define for brevity $\|\{A_i\}_{i\in[N]}\|_{2,1} = \sum_{i=1}^{n} \sum_{j=1}^{n} \|[(A_1)_{ij}, (A_2)_{ij}, \dots, (A_N)_{ij}]\|_2$. With this notation in place, we first state the main assumptions and outline the compatibility condition associated with the group sparse model [30, Chapter 8].

Assumption 1: The disturbances $\{w_i(k)\}$ are i.i.d. Gaussian random variables $\mathcal{N}(0, \sigma^2)$. □

Definition 1 ([30]): Let $S \subset [n] \times [n]$ and $S^c := ([n] \times [n]) \setminus S$. We say that the *compatibility condition* holds for the index set S if for any $\{A_i \in \mathbb{R}^{n\times n}\}_{i\in[N]}$ with $\|\{A_i^{S^c}\}_{i\in[N]}\|_{2,1} \leq 3\|\{A_i^S\}_{i\in[N]}\|_{2,1}$, it holds that

$$\|\{A_i^S\}_{i\in[N]}\|_{2,1}^2 \leq \frac{|S| \sum_{i=1}^{N} \sum_{\tau=1}^{P} \|A_i x_i(\tau)\|_2^2}{P\phi(S)} \quad (7)$$

for some constant $\phi(S) > 0$. Moreover, we define as $\mathcal{S}$ the collection of sets S for which the compatibility condition holds. □

We note that, when S coincides with the support of the matrices $\{A_i\}_{i\in[n]}$, this technical condition provides bounds on the values of the non-zero entries of the matrices.

Hereafter, we let $\{A_i^{MT}\}_{i\in[N]}$ be an optimal solution of the system identification problem (4), and we denote as $\{A_i^\star\}_{i\in[N]}$ the *true* matrices of the LTI systems in (1). The following theorem provides an error bound for the proposed multi-task system identification (4) relative to the (true) system matrices $\{A_i^\star\}_{i\in[N]}$.

Theorem 1: Let $\{A_i^\star\}_{i\in[N]}$ be the true matrices of the LTI systems and define $S^\star = \{(i,j) \mid [(A_1^\star)_{ij}, (A_2^\star)_{ij}, \ldots, (A_N^\star)_{ij}] \neq 0\}$. Let Assumption 1 hold and suppose that $S^\star \in \mathcal{S}$. Define

$$\lambda_0 := \frac{2\sqrt{M}}{nP}\left(1+\sqrt{\frac{4\gamma+8\log n}{N}}+\frac{4\gamma+8\log n}{N}\right)^{\frac{1}{2}} \quad (8)$$

for some $\gamma > 0$, where $M = \sigma^2 \max\limits_{1\leq i\leq N, 1\leq j\leq n} \sum_{k=1}^{P}[(x_i^{(j)})_k]^2$. If $\lambda \geq 4nNP\lambda_0$, then, the following bound holds

$$\sum_{i=1}^{N}\sum_{\tau=1}^{P}\|(A_i^\star - A_i^{MT})x_i(\tau)\|_2^2 \leq \frac{24\lambda^2|S^\star|}{PN\phi(S^\star)} \quad (9)$$

with probability at least $1-\mathrm{e}^{-\gamma}$. □

From Theorem 1, it can be seen that the average error $E(N,P) := \frac{1}{PN}\sum_{i=1}^{N}\sum_{i=1}^{N}\sum_{\tau=1}^{P}\|(A_i^\star - A_i^{MT})x_i(\tau)\|_2^2$ is of the order

$$E(N,P) \approx O\left(\frac{|S^\star|}{\phi(S^\star)}P^{-1}(1+N^{-\frac{1}{2}}+N^{-1})\right),$$

when taking $\lambda = O(\sqrt{P}\sqrt{N+\sqrt{N}+1})$. Interestingly, by increasing the number of LTI systems in the multi-task system identification (i.e. N increases), the average error $E(N,P)$ decreases; however, when $N\to\infty$, the error does not tend to 0. This can be understood as a plateau in the ability to "transfer information" between systems.

One possible shortcoming of Theorem 1 is that the set $S^\star$ describing the (common) support of the matrices $\{A_i^\star\}_{i\in[N]}$ is assumed to satisfy the compatibility condition. In the following, we offer an additional error bound for cases where $\{A_i^\star\}_{i\in[N]}$ is not guaranteed to satisfy the compatibility condition; the error bound leverages the notion of oracle [30].

Theorem 2: Consider the *oracle* $\{A_i^\dagger\}_{i\in[N]}$ defined as:

$$\{A_i^\dagger\}_{i\in[N]} \in \arg\min_{\{A_i\}:S_{\{A_i\}}\in\mathcal{S}}\left\{\sum_{i=1}^{N}\sum_{\tau=1}^{P}\|(A_i^\star - A_i)x_i(\tau)\|_2^2 + \frac{4\lambda^2|S_{\{A_i\}}|}{PN\phi\left(S_{\{A_i\}}\right)}\right\}, (10)$$

where $S_{\{A_i\}} = \{(i,j) \mid [(A_1)_{ij}, (A_2)_{ij}, \ldots, (A_N)_{ij}] \neq 0\}$, and let Assumption 1 hold. Set λ_0 and λ as in Theorem 1. Then, the following bound holds with probability at least $1-\mathrm{e}^{-\gamma}$, for a given $\gamma > 0$:

$$\sum_{i=1}^{N}\sum_{\tau=1}^{P}\|(A_i^\star - A_i^{MT})x_i(\tau)\|_2^2 \leq 6\sum_{i=1}^{N}\sum_{\tau=1}^{P}\|(A_i^\star - A_i^\dagger)x_i(\tau)\|_2^2 + \frac{24\lambda^2|S^\dagger|}{PN\phi(S^\dagger)}, \quad (11)$$

where $S^\dagger = \{(i,j) \mid [(A_1^\dagger)_{ij}, (A_2^\dagger)_{ij}, \ldots, (A_N^\dagger)_{ij}] \neq 0\}$. □

Theorem 2 asserts that the error incurred by the multi-task system identification (4) is bounded by the estimation error associated with the oracle plus an additional term modeling the error between the estimated matrices and the oracle. Here, the oracle represents the best achievable estimation when considering matrices with support index set that satisfies the compatibility condition.

IV. Handling Systems with Other Similarities

In this section, we explain how the proposed multi-task system identification method can be adapted to LTI systems that feature additional similarities, as previously explained in Section II.

A. Small Heterogeneity

Consider the case where, for any pair A_i, A_j, $i,j\in[N]$, there exists $\epsilon > 0$ such $\|A_i - A_j\|_F^2 \leq \epsilon$. This case is referred to as "small heterogeneity" in [23]. With this prior, the multi-task system identification problem can be formulated as

$$\min_{\{A_k\}_{k\in[N]}} \sum_{i=1}^{N}\mathcal{L}_i(A_i) + \lambda\sum_{i=1}^{N}\sum_{j=i}^{N}\|A_i - A_j\|_F^2, \quad (12)$$

where we recall that $\mathcal{L}_i(A_i)$ is the LS fit for the ith system and where the regularization term penalizes large deviations between the estimated matrices [20].

Problem (12) is convex and can be solved in closed form. However, to avoid computationally-heavy matrix inversions (especially when the dimension b is large and several systems are considered in the system identification process), Algorithm 1 can be utilized to solve (12) by replacing [S2.2] with the following n^2 parallel computations:

$$y_{ij} = \frac{z_{ij} + 2\alpha\lambda s_{ij}[1,1,\cdots,1]}{2\alpha\lambda N + 1}, \quad i,j\in[N], \quad (13)$$

where $s_{ij} = \sum_{\ell=1}^{N}(Z_\ell)_{ij}$.

From the update (13), it can be seen that $y_{ij} \to (s_{ij}/N)[1,1,\cdots,1]$ as $\lambda\to\infty$, thus setting all the system matrices to be the same. On the other hand, by setting $\lambda = 0$ one recovers the LS method. Even in this case, cross-validation procedures can be utilized to find the value of λ such that the estimated matrices yield the lowest error on test data [21].

The estimation performance of (12) can be analyzed by deriving bounds between an optimal solution of (12) and the one of the LS method. Since the bound is straightforward, and because of space limitations, we omit this result from the paper.

B. Linear Combinations

Lastly, we comment on an additional "similarity" where the system matrices are (approximately) linearly dependent. Precisely, we consider a scenario where for a subset of systems, there exists coefficients $\{\alpha_{i,j}\in\mathbb{R}\}$ with $\alpha_{i,j}\neq 0$ for some $j\in[N]$ such that $A_i \approx \sum_{j=1,j\neq i}^{N}\alpha_{ij}A_j$.

To formalize the setup, suppose that $q \ll N$ of the matrices $\{A_i\}_{i\in[N]}$ are such that the remaining $N-q$ can be represented as a linear combination of these q matrices. Consider then building the $n^2\times N$ matrix $[\mathrm{vec}(A_1), \mathrm{vec}(A_2), \ldots, \mathrm{vec}(A_N)]$; it follows that this matrix has rank $q \ll N$. Based on this observation, we propose to formulate a multi-task system identification problem for this case as

$$\min_{\{A_k\}_{k\in[N]}} \sum_{i=1}^{N} \mathcal{L}_i(A_i) \\ + \lambda \left\| [\mathrm{vec}(A_1), \mathrm{vec}(A_2), \dots, \mathrm{vec}(A_N)] \right\|_* ,$$

where the regularization function promotes sparsity in the singular values of the matrix $[\mathrm{vec}(A_1), \mathrm{vec}(A_2), \dots, \mathrm{vec}(A_N)]$ (see, e.g., [31], [32]).

Problem (14) is convex and with a composite cost where the regularization function is not differentiable [31], [32]. Still, the proximal-gradient algorithm tabulated as Algorithm 1 can be modified to solve (14). In particular, one can replace the step [S2.2] is Algorithm 1 with $\mathrm{prox}_{\alpha\lambda\|[\mathrm{vec}(A_1),\mathrm{vec}(A_2),\dots,\mathrm{vec}(A_N)]\|_*}(\{Z_i\}_{i\in[N]})$; this proximal map affords a closed-form solution given by:

$$\bar{Y} = U\mathrm{diag}(\{\max\{\sigma_i - \alpha\lambda, 0\}\})V^*, \qquad (14)$$

where the singular value decomposition of $[\mathrm{vec}(A_1), \mathrm{vec}(A_2), \dots, \mathrm{vec}(A_N)]$ is $U\mathrm{diag}(\{\sigma_i\})V$. The matrices $\{Y_\ell\}_{\ell\in[N]}$ are then extracted from the columns of $\bar{Y}$. From the computation of $\bar{Y}$, it can be seen that higher values of λ lead to a higher number of singular values of $\bar{Y}$ that are set to zero.

While the effectiveness of nuclear norm minimization approaches has been verified numerically in the literature, identifiability results and analytical bounds for the estimation error are available only under given assumptions on the regressors [31], [32] that may not be applicable to (14); see also [33]. Deriving analytical error bounds for (14) is subject of our ongoing investigations.

Remark 1: The proposed multi-task system identification problems can be extended to cases where the system matrices $\{A_i\}_{i\in[N]}$ are similar according to more than one of the priors (s1)–(s2). For example, if the matrices have a common sparsity pattern and the differences in the non-zero entries are small, one can utilize the composite regularization function $\lambda_1 \sum_{i=1}^{N}\sum_{j=i}^{N} \|A_i - A_j\|_F^2$ $+\lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{N} \|[(A_1)_{ij}, (A_2)_{ij}, \dots, (A_N)_{ij}]^\top\|_2$, where $\lambda_1, \lambda_2 \geq 0$ are tuning parameters. □

V. Numerical Experiments

A. Experiments on brain networks

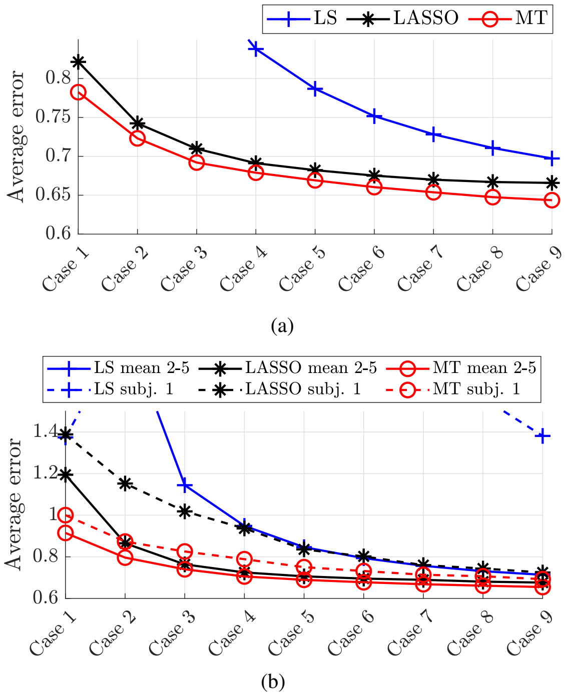
We test the proposed method for the problem of estimating the dynamics of brain networks, using data corresponding to the resting state fMRI from the Human Connectome Project (HCP)[3] [34], [35], [41]. Here, $x_i(t)$ is a 116-dimensional blood-oxygen-level-dependent (BOLD) time series for 116 parcellations of the brain of the i-th subject. Our goal here is to estimate $N = 5$ dynamical systems of the form $x_i(t+1) = A_i x_i(t) + w_i(t)$, that model the evolution of BOLD signal when the individual is in a resting state, with $w_i(t)$ capturing process noise (the model does not contain external inputs u_i due to the resting state condition).

Since the matrices $\{A_i\}_{i\in[5]}$ are unknown, we consider the following error for each system:

$$\mathcal{E}(A) := \frac{1}{n}\sum_{k=1}^{n} \frac{\sum_{i=1}^{p}(x_i(k) - [Ax_i](k))^2}{\sum_{i=1}^{p}(x_i(k) - \bar{x}(k))^2},$$

where n is the length of the testing vector, p is the number of testing data and $\bar{x}(k) := \frac{1}{p}\sum_{i=1}^{p} x_i(k)$. Note that $1 - \mathcal{E}(A)$ is precisely the average R^2 indicator of [35].

We consider three different methods: (i) the LS estimator (2), which is utilized per individual; (ii) the Least Absolute Shrinkage and Selection Operator (LASSO), which is again utilized per individual as proposed in [35]; and, (iii) the proposed method (3) with the group-sparsity regularization function (4). The rationale behind the group-sparsity is that the brain dynamics should exhibit the same effective connectivity between parcellations, though the remaining entries acknowledge the diversity in intensities of the interactions across individuals. We note that the effectiveness of the LS and LASSO has been experimentally validated in [35], where their estimation accuracy has been compared with several identification methods. Moreover, we performed a cross-validation procedure to optimize the performance of the LASSO.



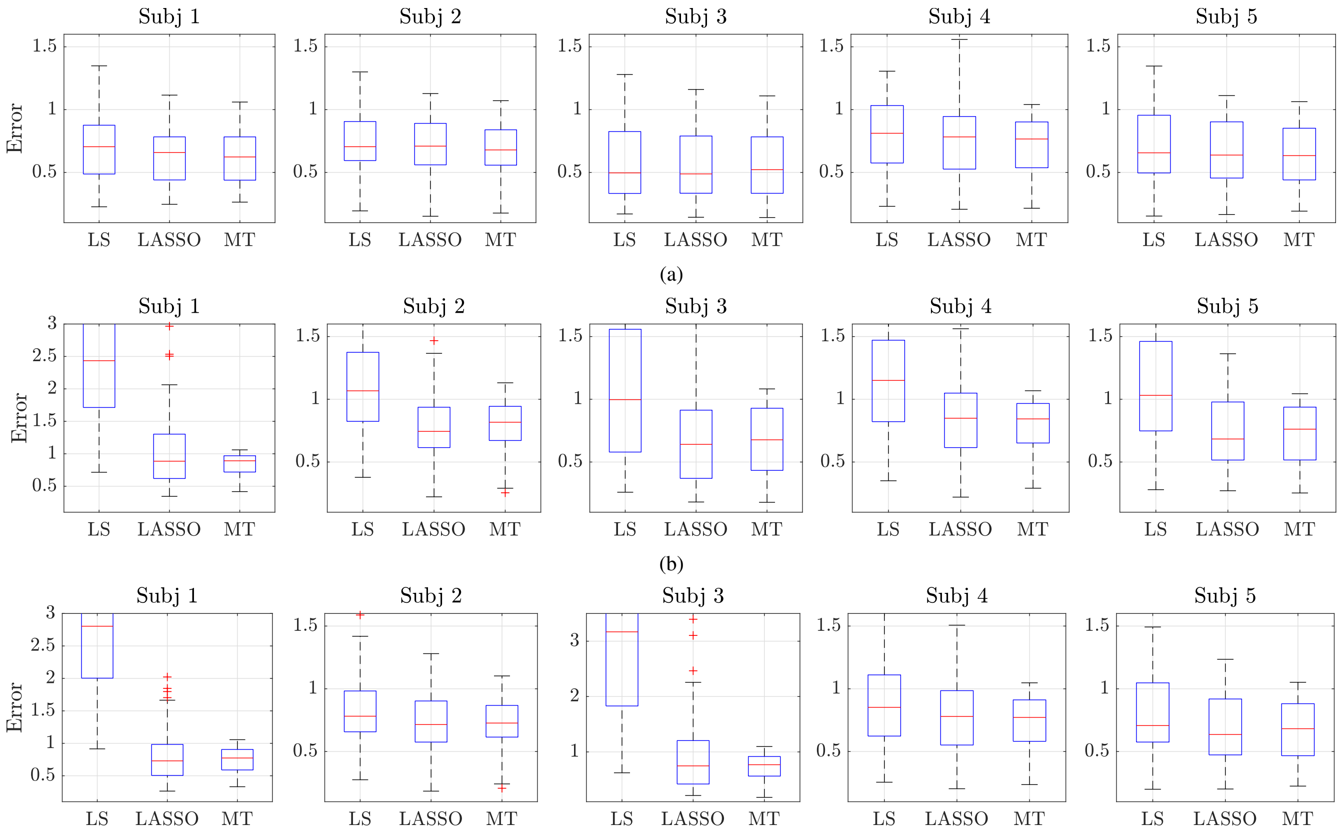
(a)

(b)

Fig. 2. (a) Mean error of LS, LASSO and multi-task (MT) system identification; "Case k" means that $100k$ training data points are available for each subject ($k = 1, 2, \cdots, 9$). (b) Mean error for subjects 2-5 and error for subject 1. "Case k" means that $25k$ fMRI scans are used for subjects 1 (dashed line) while $100k$ (solid line) scans are used for subjects 2-5.

In Figure 2, we compare the LS, LASSO and our approach (which is labeled as "MT") in two cases: (a) the same amount of training data is utilized for the five subjects; and, (b) for subject 1, we utilize only 25% of the training data points with respect to the other subjects 2-5. We use 100 test points. In Figure 2(a) we plot the mean error across the

[3] Data available at https://wiki.humanconnectome.org/



(a)

(b)

(c)

Fig. 3. Comparison between LS, LASSO and multi-task (MT) system identification (a) Case 1: 900 training data points for each subject. (b) Case 2: For subject 1, 75 training data points, and 300 for subjects 2-5. (c) Case 3: For subject 1 and 3, 150 training data points, and 600 for subjects 2, 4, and 5. In the box plots, the red center line, box limits, and whiskers represent the median, upper and lower quartiles, and the smallest and largest samples, respectively. Red crosses indicate outliers.

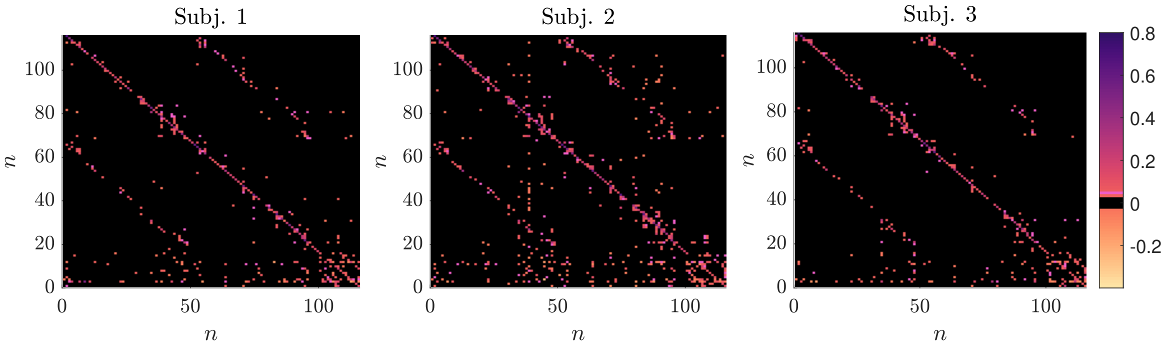


Fig. 4. Estimated matrix $\hat{A}_i$ for Case 3, individuals 1, 2 and 3, for $n = 116$ brain parcellations.

subjects 1-5; in Figure 2(b) we plot the mean error across the subjects 2-5 and the error for subject 1, for which fewer fMRI readings are available. The proposed method outperforms the LS and the LASSO, on par with the number of fMRI scans in both cases. The merits of the proposed method are particularly evident in Figure 2(b), where the proposed method significantly outperforms the LASSO for the subject 1; on the other hand, the LS is ill-conditioned and does not return meaningful estimates. This shows the ability to leverage information and data (in this case, fMRI readings) from the dynamics of subjects 2-5 to assist the estimation of the dynamics in subject 1.

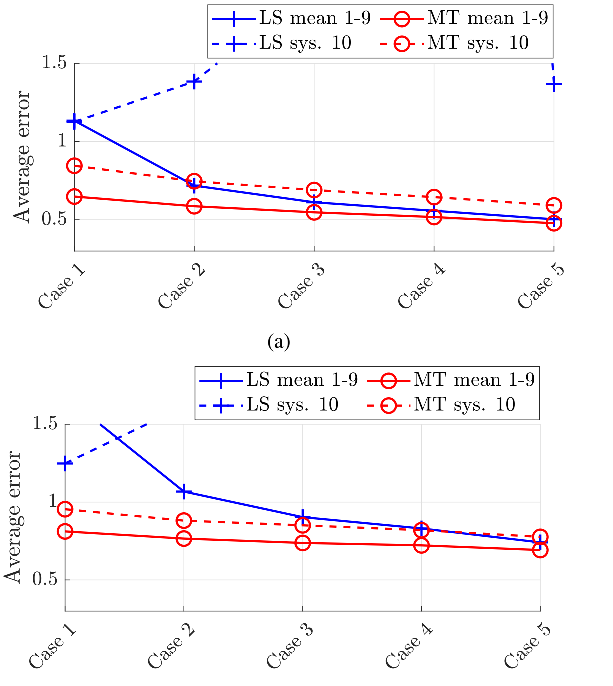
To provide additional comparisons other than the mean error, Figure 3 shows the box plots for the LS, the LASSO, and the proposed approach in three different scenarios. In particular, Figure 3(a) shows that proposed multi-task identification method can achieve a smaller or comparable error (on average) than LS or LASSO when trajectories of 900 time steps are used for each subject (and these training trajectories are sufficiently rich). Figure 3(b) considers the case where 75 training data points are available for subject 1 and 300 for subjects 2-5. Here, the LS does not perform well due to ill-conditioning. The performance of the LASSO is comparable with the one of the proposed method in terms of median; however, the proposed method shows smaller upper and lower quartiles. Moreover, Figure 3(c) considers the case where fewer fMRI readings are available for subjects 1 and 3; the proposed method performs better than the LASSO in terms of quartiles and has a significantly less error deviation across the parcellations.

Finally, a representative example of the estimated matrices $\hat{A}_i$ for the subjects 1-3 is provided in Figure 4. The estimated matrices are the ones obtained in the case considered in Figure 3(c), where subjects 1 and 3 have fewer training points. It is possible to notice that the three matrices have zeros in many common entries. Based on this result, we will explore additional regularization methods that will combine group sparsity with (entry-wise) sparsity.

B. Experiments on synthetic data

We provide additional results on synthetic data. We consider 10 systems as in (1), where $\{A_i\}_{i\in[10]} \in \mathbb{R}^{50\times 50}$, $\{B_i\}_{i\in[10]} \in \mathbb{R}^{50\times 4}$, $u_i(t)$ is the vector of all ones in $\mathbb{R}^4$, i.e. $u_i(t)$ is constant vector and $w_i(t) \sim \mathcal{N}(0, 0.1^2)$. We consider two different cases: common sparsity and linear combinations. We compare the LS estimator (2) and the proposed method (3) with the group-sparsity regularization (4) and nuclear norm regularization (14).

Figure 5 compares the LS and our approach in two cases: (a) all the 10 systems can be represented by a linear combination of 3 systems and only 25% of the training data points are accessible for the tenth system with respect to the other systems 1-9; (b) all the 10 systems have the same sparsity pattern and only 25% of the training data points are accessible for the tenth system with respect to the other systems 1-9. The testing is on 60 data points. In Figure 5(a), we plot the mean error across systems 1-9 as well as the error for system 10. The proposed method outperforms the LS approach in both the mean error and the error for system 10, especially in the case of only a small number of data available. In Figure 5(b), we can observe similar results.



(a)

(b)

Fig. 5. Mean error curve to compare LS and multi-task (MT) system identification methods. (a) Linear combinations. (b) Common sparsity. "Case k" means that $10k + 10$ samples of the trajectory are used for system 10 (dash line) while $40k + 40k$ (solid line) are used for system 1-9, $k = 1, 2, 3, 4$. In "Case 5", 75 data points are used for system 10 (dashed line) while $300k$ (solid line) are used for system 1-9.

REFERENCES

[1] L. Ljung, *System Identification: Theory for the user*. Prentice-Hall, 1987.

[2] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung, "Regularized system identification: Learning dynamic models from data," 2022.

[3] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 909–924, 2019.

[4] J. Coulson, J. Lygeros, and F. Dörfler, "Data-enabled predictive control: In the shallows of the DeePC," in *European Control Conference*, 2019, pp. 307–312.

[5] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 269–296, 2020.

[6] J. Berberich, J. Köhler, M. A. Müller, and F. Allgöwer, "Data-driven model predictive control with stability and robustness guarantees," *IEEE Transactions on Automatic Control*, vol. 66, no. 4, pp. 1702–1717, 2020.

[7] V. Krishnan and F. Pasqualetti, "On direct vs indirect data-driven predictive control," in *IEEE Conference on Decision and Control*, 2021, pp. 736–741.

[8] L. Li, C. De Persis, P. Tesi, and N. Monshizadeh, "Data-based transfer stabilization in linear systems," *arXiv preprint arXiv:2211.05536*, 2022.

[9] J. C. Willems, P. Rapisarda, I. Markovsky, and B. L. De Moor, "A note on persistency of excitation," *Systems & Control Letters*, vol. 54, no. 4, pp. 325–329, 2005.

[10] C. De Persis and P. Tesi, "On persistency of excitation and formulas for data-driven control," in *IEEE Conference on Decision and Control*, 2019, pp. 873–878.

[11] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *International Conference on Machine Learning*, 2019, pp. 5610–5618.

[12] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*, 2018, pp. 439–473.

[13] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[14] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *American control conference*, 2019, pp. 5655–5661.

[15] Y. Zheng and N. Li, "Non-asymptotic identification of linear dynamical systems using multiple trajectories," *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1693–1698, 2020.

[16] L. Xin, G. Chiu, and S. Sundaram, "Learning the dynamics of autonomous linear systems from multiple trajectories," in *American Control Conference*, 2022.

[17] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto, "System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 2933–2945, 2014.

[18] A. Chiuso and G. Pillonetto, "System identification: A machine learning perspective," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 281–304, 2019.

[19] Y. Sun, S. Oymak, and M. Fazel, "Finite sample system identification: Optimal rates and the role of regularization," in *Learning for Dynamics and Control*, 2020, pp. 16–25.

[20] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[21] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical learning with sparsity," *Monographs on statistics and applied probability*, vol. 143, p. 143, 2015.

[22] L. Xin, L. Ye, G. Chiu, and S. Sundaram, "Identifying the dynamics of a system by leveraging data from similar systems," in *American Control Conference*, 2022.

[23] H. Wang, L. F. Toso, and J. Anderson, "Fedsysid: A federated approach to sample-efficient system identification," *arXiv preprint arXiv:2211.14393*, 2022.

[24] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 109–117.

[25] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[26] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[27] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[28] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[29] J. Huang and T. Zhang, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[30] P. Bhlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, 2011.

[31] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Sparse and low-rank matrix decompositions," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 1493–1498, 2009.

[32] M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *IEEE Transactions on Signal Processing*, vol. 63, no. 10, pp. 2663–2677, 2015.

[33] K. Mohan and M. Fazel, "Reweighted nuclear norm minimization with application to system identification," in *Proceedings of the 2010 American Control Conference*. IEEE, 2010, pp. 2953–2959.

[34] P. Srivastava, E. Nozari, J. Z. Kim, H. Ju, D. Zhou, C. Becker, F. Pasqualetti, G. J. Pappas, and D. S. Bassett, "Models of communication and control for brain networks: distinctions, convergence, and future outlook," *Network Neuroscience*, vol. 4, no. 4, pp. 1122–1159, 2020.

[35] E. Nozari, M. A. Bertolero, J. Stiso, L. Caciagli, E. J. Cornblath, X. He, A. S. Mahadevan, G. J. Pappas, and D. S. Bassett, "Is the brain macroscopically linear? a system identification of resting state dynamics," *arXiv preprint arXiv:2012.12351*, 2020.

[36] B. Turan and M. Alizadeh, "Competition in electric autonomous mobility on demand systems," *IEEE Transactions on Control of Network Systems*, 2021.

[37] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[38] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery," *Convex optimization in signal processing and communications*, pp. 42–88, 2009.

[39] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[40] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[41] S. Gu, F. Pasqualetti, M. Cieslak, Q. K. Telesford, A. B. Yu, A. E. Kahn, J. D. Medaglia, J. M. Vettel, M. B. Miller, S. T. Grafton *et al.*, "Controllability of structural brain networks," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

Acknowledgements

The authors would like to thank Dr. Erfan Nozari (University of California at Riverside) for the assistance with the data used in the simulations, and Dr. Stephen Becker (University of Colorado Boulder) for the discussion on the group Lasso.

Appendix

Proofs of Theorems 1 and 2. We provide a sketch of the proofs of Theorems 1 and 2. To this end, we introduce some additional notation. For the ith system, collect the recorded trajectories in the matrix $X_i := [x_i(1), ..., x_i(P_i)]$; then, the LS fit $\mathcal{L}_i(A_i)$ for the ith system can be equivalently expressed as

$$\mathcal{L}_i(A_i) = \|\tilde{\mathbf{y}}_i - \tilde{X}_i \tilde{\mathbf{a}}_i\|_2^2,$$

where $\tilde{\mathbf{y}}_i := \mathrm{vec}([x_i(2) - B_i u_i(1), ..., x_i(P_i+1) - B_i u_i(P_i)])$, $\tilde{\mathbf{a}}_i = \mathrm{vec}(A_i)$ and $\tilde{X}_i = X_i^\top \otimes I_n$. Moreover, let $\tilde{\mathbf{A}} := [\tilde{\mathbf{a}}_1, ..., \tilde{\mathbf{a}}_N]$ collect all the vectorized system matrices $\{A_i\}_{i\in[N]}$, $\mathbf{x}_i^{(j)}$ the jth column of $\tilde{X}_i$, and denote as $\|\tilde{\mathbf{A}}\|_{2,1}$ the sum of the l_2-norm of each row of $\tilde{\mathbf{A}}$. Finally, recall that for any matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{n^2\times N}$, $\tilde{\mathbf{A}}_S$ is a matrix formed by selecting the rows of $\tilde{\mathbf{A}}$ indexed by S and setting to zero the other rows.

With this notation in place, and recalling that $\{A_i^{MT}\}_{i\in[N]}$ denotes an optimal solution of (4), we have the following result.

Lemma 1: Let Assumption 1 hold. Let λ_0 be defined as in Theorem 1, and take $\lambda \geq 4nNP\lambda_0$. Then, for any $\tilde{\mathbf{A}} \in \mathbb{R}^{n^2\times N}$ and any $S \in \mathcal{S}$ satisfying the compatibility condition, the following bound holds with probability at least $1 - \mathrm{e}^{-\gamma}$:

$$\begin{aligned}\sum_{i=1}^{N} \|\tilde{X}_i\tilde{\mathbf{a}}_i^\star - \tilde{X}_i\tilde{\mathbf{a}}_i^{MT}\|_2^2 + \frac{\lambda}{\sqrt{N}}\|\tilde{\mathbf{A}}^{MT} - \tilde{\mathbf{A}}_S\|_{2,1} \\ \leq 6\sum_{i=1}^{N}\left\|\tilde{X}_i\tilde{\mathbf{a}}_i^\star - \tilde{X}_i\tilde{\mathbf{a}}_{S,i}\right\|_2^2 + \frac{24\lambda^2|S|}{PN\phi^2(S)} \quad (15)\end{aligned}$$

for any given $\gamma > 0$. □

The proof of Lemma 1 is omitted due to space limitations. Concisely, the proof of Lemma 1 leverages some arguments from Chapters 6 and 8 in [30]. First, we establish basic inequalities for our problem similarly to [30, Chapter 6], and bound the empirical process by [30, Lemma. 8.5]. Then, we derive an inequality similar to [30, Lemma. 6.3] for our group sparse model. Finally, combining these inequalities, we prove the inequalities in Lemma 1 (similar to [30, Thm. 6.2]). The full proof will be made available on an extended version online.

Based on the result of Lemma 1, the bound in Theorem 1 can be shown by setting $\tilde{\mathbf{A}}_S = \tilde{\mathbf{A}}^\star$ in (15), thus obtaining

$$\begin{aligned}\sum_{i=1}^{N} \|\tilde{X}_i\tilde{\mathbf{a}}_i^\star - \tilde{X}_i\tilde{\mathbf{a}}_i^{MT}\|_2^2 + \frac{\lambda}{\sqrt{N}}\|\tilde{\mathbf{A}}^{MT} - \tilde{\mathbf{A}}^\star\|_{2,1} \\ \leq \frac{24\lambda^2|S^\star|}{PN\phi^2(S^\star)} \quad (16)\end{aligned}$$

and noticing that $\sum_{i=1}^{N}\|\tilde{X}_i\tilde{\mathbf{a}}_i^\star - \tilde{X}_i\tilde{\mathbf{a}}_i^{MT}\|_2^2 = \sum_{i=1}^{N}\sum_{\tau=1}^{P}\|(A_i^\star - A_i^{MT})x_i(\tau)\|_2^2$. On the other hand, the bound in Theorem 2 can be shown by setting $\tilde{\mathbf{A}}_S = \tilde{\mathbf{A}}^\dagger$ in (15), where we recall that $\tilde{\mathbf{A}}^\dagger$ is the oracle solution.