Joint Network Topology Inference in the Presence of Hidden Nodes

Madeline Navarro, *Student Member, IEEE*, Samuel Rey, *Member, IEEE*, Andrei Buciulea, *Student Member, IEEE*, Antonio G. Marques, *Senior Member, IEEE*, and Santiago Segarra, *Senior Member, IEEE*

***Abstract*—We investigate the increasingly prominent task of jointly inferring *multiple* networks from nodal observations. While most *joint* inference methods assume that observations are available at all nodes, we consider the realistic and more difficult scenario where a subset of nodes are *hidden* and cannot be measured. Under the assumptions that the partially observed nodal signals are graph stationary and the networks have similar connectivity patterns, we derive structural characteristics of the connectivity between hidden and observed nodes. This allows us to formulate an optimization problem for estimating networks while accounting for the influence of hidden nodes. We identify conditions under which a convex relaxation yields the sparsest solution, and we formalize the performance of our proposed optimization problem with respect to the effect of the hidden nodes. Finally, synthetic and real-world simulations provide evaluations of our method in comparison with other baselines.**

***Index Terms*—Graph learning, network topology inference, hidden nodes, graph signal processing, graph stationarity, multi-layer graphs.**

I. Introduction

In recent years, graphs have become a staple model of the irregular (non-Euclidean) structure commonly found in contemporary data. Disciplines like signal processing often rely on graphs to capture the underlying irregular domain of the signals, where such successful applications include genetics, brain networks, and communications [2]–[4]. Nevertheless, despite the popularity of graph-based methods, in practice the topology of the graph is often not readily available, spurring the development of graph learning algorithms [5]–[7] to infer the network topology from a set of nodal observations.

Indeed, the task of *network topology inference*, also known as *graph learning*, has emerged as a vibrant research area within graph signal processing (GSP) [8]–[11]. A crucial assumption for learning the graph topology is the statistical relationship between the signals and the unknown topology. Different assumptions lead to different methods, with noteworthy examples including correlation networks and (Gaussian) Markov random fields ((G)MFR) [2], [5], [12], smooth (local total variation) models [13]–[15], GSP-based approaches [16]–[18], and models with more elaborate graph priors [19], [20]. A common feature of the previous works is that they focus on learning a single graph. However, many contemporary setups involve *multiple related networks*, each with a subset of signals. Some examples include brain analytics, where observations from different *patients* are used to estimate their brain functional networks; social networks, where the same set of users may present different types of *interactions*; or multi-hop communication networks in dynamic environments, where a network needs to be inferred for each *time instant*. Intuitively, in situations where several closely related networks exist, approaching the problem in a joint fashion can boost the performance of network topology inference by harnessing the relationships among graphs [21]–[26].

Despite the clear benefits, joint network topology inference approaches usually assume that observations from every node are available, which is often not the case. In many relevant scenarios, the observed signals correspond only to a subset of the nodes in the whole graph, while the remaining nodes stay unobserved or *hidden*. Ignoring the presence of the hidden nodes can drastically hinder the performance of the graph learning algorithms. Nevertheless, accounting for their influence is not a trivial endeavor since the inference task becomes ill-posed. For *single network* inference, some works dealing with this challenging setting include graphical models [27], [28], inference of linear Bayesian networks [29], nonlinear regression [30], and stationary-based algorithms [31], [32]. However, the presence of hidden nodes is yet to be addressed for several unknown graphs. Since the key to joint topology inference is exploiting the similarity of the graphs, it is crucial to model the influence of the hidden nodes to measure the graph similarity between nodes that remain unobserved.

To this end, we propose a topology inference method that simultaneously performs *joint estimation of multiple graphs* and *accounts for the presence of hidden variables*. Under the assumption that the observed signals are realizations of a random process that is *stationary* on the graph [10], [33], we formalize the relationship between the nodal observations and the unknown networks under the influence of the hidden nodes. The joint formulation necessitates exploiting graph similarities, not only with respect to observed nodes but also to hidden ones. To accomplish this, we carefully model the structure associated with latent variables and exploit it with

This work was partially supported by NSF under award CCF-2008555, Spanish Fed. Grants FPU17-04520 and SPGraph PID2019-105032GB-I00, and URJC grant F861. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-17-S-0002. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Army or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Preliminary results were presented at ICASSP 2022 [1]. *(Corresponding author: M. Navarro)*

M. Navarro and S. Segarra are with the Dept. of ECE, Rice University, Houston, TX 77005 USA (e-mail: {nav,segarra}@rice.edu). S. Rey, A. Buciulea, and A. G. Marques are with the Dept. of Signal Theory and Comms., King Juan Carlos University, 28933 Madrid, Spain (e-mail: {samuel.rey.escudero, andrei.buciulea, antonio.garcia.marques}@urjc.es).

a regularization inspired by the group Lasso penalty [34]. Finally, we conduct thorough mathematical and numerical analyses of the proposed approach, where we show the conditions under which it recovers the sparsest solution and bounds the error of the estimated graphs, and we evaluate its performance and the hidden variables' detrimental influence through simulations with synthetic and real-world data.

Related work and contributions. Early methods for joint graph learning were introduced in [22] assuming that observations follow a GMRF and, later on, in [23] followed by a joint inference method for graph stationary signals. However, both works assumed that observations from the whole graphs were available. At the same time, the influence of hidden nodes when learning a single graph was studied in [27] and [32] assuming that the observations adhered respectively to a GMRF or a graph-stationary model. On the other hand, the relevant task of learning several graphs in the presence of hidden nodes has only been considered under GMRF assumptions in the preliminary results from [35]. In contrast, in this paper, we (i) build over our previous work from [1] for joint graph learning with hidden variables under the more lenient assumption of stationary observations; and (ii) develop a theoretical analysis to characterize how the hidden nodes influence the quality of the estimated graphs. Finally, note that GMRF and graph stationarity are intrinsically different models for the observations, resulting in materially different inference algorithms and, even more relevant for the problem at hand, requiring different methods to encourage graph similarities with respect to both observed and hidden nodes.

To summarize, our main contributions are:

- We design a convex optimization problem to jointly learn the topology of several related graphs in the presence of hidden variables under graph-stationary observations.
- We rely on a regularization inspired by group Lasso to model the similarity between hidden nodes and hence harness the similarity of the entire node set, both hidden and observed nodes.
- We derive theoretical guarantees for the recoverability of the estimated graphs in the presence of hidden nodes.
- We evaluate the performance of the proposed approach and compare it with state-of-the-art alternatives in synthetic and real-world datasets.

The remainder of the paper is organized as follows. Section II introduces GSP concepts necessary for our proposed network topology inference method and its theoretical guarantees. We introduce in Section III the task of learning graphs in the presence of hidden nodes. In Section IV we present our proposed optimization problem that accounts for hidden nodes, along with its convex relaxation. We provide theoretical guarantees for the viability and performance of our method in Section V, which are validated by several synthetic and real-world experiments in Section VI. Finally, a concluding discussion is provided in Section VII.

II. Fundamentals of GSP

We introduce notation and concepts in GSP to characterize the statistical relationship between the network topology and measurements on nodes, both observed and hidden.

Notation. For a matrix $\mathbf{Y} \in \mathbb{R}^{M\times N}$, $\text{vec}(\mathbf{Y}) \in \mathbb{R}^{MN}$ denotes the vertical concatenation of the columns of $\mathbf{Y}$. We let calligraphic letters denote index sets, where, given any matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$ and any vector $\mathbf{x} \in \mathbb{R}^N$, we let $\mathbf{X}_{\mathcal{C},\cdot}$ and $\mathbf{X}_{\cdot,\mathcal{C}}$ respectively return the rows and columns of $\mathbf{X}$ selected from index set $\mathcal{C}$ and $\mathbf{x}_{\mathcal{C}}$ returns the entries of $\mathbf{x}$ selected from $\mathcal{C}$. The notation $\mathbf{I}_M$ denotes the identity matrix of size $M\times M$, while $\mathbf{1}_{M\times N}$ and $\mathbf{0}_{M\times N}$ respectively represent matrices of all ones and zeros of size $M \times N$. We let $\mathcal{D}$, $\mathcal{L}$, and $\mathcal{U}$ respectively denote the indices of the diagonal, lower triangular, and upper triangular entries of a vectorized square matrix, i.e., for any matrix $\mathbf{Y} \in \mathbb{R}^{M\times M}$ and $\mathbf{y} = \text{vec}(\mathbf{Y})$, we have that $\mathbf{y}_{\mathcal{D}}$ contains the diagonal entries of $\mathbf{Y}$. We define $\mathbf{y}_{\mathcal{L}}$ and $\mathbf{y}_{\mathcal{U}}$ similarly. The notation $O(\cdot)$ and $o(\cdot)$ denote the usual asymptotic meaning, and we say that $f \asymp g$ if $f = O(g)$ and $g = O(f)$.

Graph signal processing and graph stationarity. We consider undirected graphs of the form $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of $|\mathcal{V}| = N$ nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set such that the unordered pair $(i, j) \in \mathcal{E}$ if and only if nodes i and j are connected. A convenient representation for the structure of a graph is its adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, where $A_{ij} = A_{ji} \neq 0$ if and only if $(i, j) \in \mathcal{E}$. We may define a more general class of matrices to encode graph structure known as the graph shift operator (GSO), of which the adjacency matrix is an example [8]–[10]. Formally, the GSO is a square matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$, where $S_{ij} \neq 0$ only if $i = j$ or $(i, j) \in \mathcal{E}$. Commonly chosen GSOs include the adjacency matrix $\mathbf{A}$ and the graph Laplacian $\mathbf{L} := \text{diag}(\mathbf{A1}) - \mathbf{A}$ [8], [10]. Because we consider undirected graphs, $\mathbf{S}$ is symmetric and thus diagonalizable.

Critical to the network inference task is the statistical relationship between nodal observations and the topology of $\mathcal{G}$. We represent real-valued observations on the nodes of $\mathcal{G}$ as graph signals $\mathbf{x} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$, where x_i denotes the signal value at the i-th node. In this work, we assume that the *observations* are realizations of a random graph signal that is *stationary on* $\mathcal{G}$ [16], [33], [36], a versatile model that has shown theoretical and practical relevance. From a mathematical point of view, a random graph signal $\mathbf{x}$ is *stationary* on its underlying graph $\mathcal{G}$ if the covariance matrix of $\mathbf{x}$, denoted as $\mathbf{C}$, can be written as a (matrix) polynomial of the GSO $\mathbf{S}$, which results in $\mathbf{C}$ and $\mathbf{S}$ having the same eigenvectors [10], [33], [37], [38]. This definition includes correlation networks, where $\mathbf{C} = \mathbf{S}$ and MRFs, where $\mathbf{C} = \mathbf{S}^{-1}$, as particular cases. From a practical (generative) point of view, stationary random graph signals are particularly suited to represent consensus dynamics, heat diffusion processes, and network processes on brain structural networks [39]–[41]. Formally, under this point of view we have that the random graph signal $\mathbf{x}$ can be modelled as $\mathbf{x} = \mathbf{Hw}$, where $\mathbf{w}$ is a stochastic zero-mean white input signal and $\mathbf{H}$ performs the diffusion process on $\mathbf{w}$ that characterizes the influence of the GSO $\mathbf{S}$ on $\mathbf{x}$. To that end, the matrix $\mathbf{H}$ is assumed to be a *linear graph filter* [9], [42], [43], a matrix polynomial of the GSO $\mathbf{H} = \sum_{l=0}^{L-1} h_l \mathbf{S}^l$ with real-valued filter coefficients $\{h_l\}_{l=0}^{L-1}$ that sufficiently

models nodal behavior for many signal processing tasks, including denoising and interpolation [10], [39], [42], [44], [45]. The structure of $\mathbf{S}$ dictates the behavior of the graph signal $\mathbf{x} = \mathbf{H}\mathbf{w}$, where we may view $\mathbf{S}^l\mathbf{w}$ as the diffusion of $\mathbf{w}$ across an l-hop neighborhood. Under the diffusion model, the signal behavior at the i-th node is encoded in the diffused signal values in an $(L-1)$-hop radius. Under this setting, the graph signals are *random* with covariance $\mathbf{C} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{H}\mathbb{E}[\mathbf{w}\mathbf{w}^\top]\mathbf{H} = \mathbf{H}^2$ due to the input $\mathbf{w}$ being white. Clearly, if $\mathbf{H}$ is a polynomial of $\mathbf{S}$, so is $\mathbf{C} = \mathbf{H}^2$, showing that both point of views are equivalent.

Finally, we note that under stationarity of $\mathbf{x}$, we have that matrices $\mathbf{S}$ and $\mathbf{C}$ commute and hence, it must hold that $\mathbf{C}\mathbf{S} = \mathbf{S}\mathbf{C}$. This is a compact and tractable way to account for the graph stationarity of the observed signals and will be later on used as a constraint in our optimization problems.

III. Inference of multilayered graphs with latent variables

Let there be a set of K undirected networks $\{\mathcal{G}^{(k)}\}_{k=1}^K$ on the same set $\mathcal{V}$ of N nodes with GSOs denoted as $\{\mathbf{S}^{*(k)}\}_{k=1}^K$. We assume that for each graph there exist a set with R_k realizations of a *stationary* graph signal collected in data matrices $\mathbf{X}^{(k)} \in \mathbb{R}^{N\times R_k}$, where the R_k columns contain the nodal observations on the k-th graph. For a signal $\mathbf{x}^{(k)}$ on the k-th graph, its covariance matrix is denoted by $\mathbf{C}^{(k)} = \mathbb{E}[\mathbf{x}^{(k)}(\mathbf{x}^{(k)})^\top]$. We further assume that for every graph we do not know the entire data matrix $\mathbf{X}^{(k)}$ but only observe signal values on a subset $\mathcal{O} \subset \mathcal{V}$ of O nodes, where $\mathcal{H} := \mathcal{V}\backslash\mathcal{O}$ denotes the set of H hidden nodes. Our goal is to *estimate the subnetwork of each network $\mathcal{G}^{(k)}$ induced by $\mathcal{O}$ from partially observed graph signals*.

Under this setting, we can now formalize the task of estimating the network structure at the node subset $\mathcal{O}$ that is encoded in the GSOs $\{\mathbf{S}^{*(k)}\}_{k=1}^K$. Without loss of generality, we partition the GSO and the covariance matrix of each network as

$$\mathbf{S}^{*(k)} = \begin{bmatrix} \mathbf{S}^{*(k)}_{\mathcal{O}} & \mathbf{S}^{*(k)}_{\mathcal{O}\mathcal{H}} \\ \mathbf{S}^{*(k)}_{\mathcal{H}\mathcal{O}} & \mathbf{S}^{*(k)}_{\mathcal{H}} \end{bmatrix}, \quad \mathbf{C}^{(k)} = \begin{bmatrix} \mathbf{C}^{(k)}_{\mathcal{O}} & \mathbf{C}^{(k)}_{\mathcal{O}\mathcal{H}} \\ \mathbf{C}^{(k)}_{\mathcal{H}\mathcal{O}} & \mathbf{C}^{(k)}_{\mathcal{H}} \end{bmatrix}, \quad (1)$$

where $\mathbf{S}^{*(k)}_{\mathcal{O}\mathcal{H}} = (\mathbf{S}^{*(k)}_{\mathcal{H}\mathcal{O}})^\top$ and $\mathbf{C}^{(k)}_{\mathcal{O}\mathcal{H}} = (\mathbf{C}^{(k)}_{\mathcal{H}\mathcal{O}})^\top$ by the symmetry of $\mathbf{S}^{*(k)}$ and $\mathbf{C}^{(k)}$. The submatrices $\mathbf{S}^{*(k)}_{\mathcal{O}} \in \mathbb{R}^{O\times O}$ and $\mathbf{S}^{*(k)}_{\mathcal{H}} \in \mathbb{R}^{H\times H}$ encode the connectivity of the subnetworks of $\mathcal{G}^{(k)}$ induced by $\mathcal{O}$ and $\mathcal{H}$, respectively, while $\mathbf{S}^{*(k)}_{\mathcal{O}\mathcal{H}} \in \mathbb{R}^{O\times H}$ represents the edges connecting observed nodes to hidden nodes. We similarly define $\mathbf{C}^{(k)}_{\mathcal{O}}$, $\mathbf{C}^{(k)}_{\mathcal{H}}$, and $\mathbf{C}^{(k)}_{\mathcal{O}\mathcal{H}}$. Given the partitions in (1), we aim to estimate the subnetworks encoded in $\{\mathbf{S}^{*(k)}_{\mathcal{O}}\}_{k=1}^K$.

We also partition each $\mathbf{X}^{(k)}$ to be conformal with $\mathbf{S}^{*(k)}$ and $\mathbf{C}^{(k)}$ as $\mathbf{X}^{(k)} = [\mathbf{X}^{(k)\top}_{\mathcal{O}}, \mathbf{X}^{(k)\top}_{\mathcal{H}}]^\top$, where $\mathbf{X}^{(k)}_{\mathcal{O}} \in \mathbb{R}^{O\times R_k}$ is the data matrix containing the partially observed graph signals and $\mathbf{X}^{(k)}_{\mathcal{H}} \in \mathbb{R}^{H\times R_k}$ remains unknown. We can thus apply the partially observed *stationary* graph signals $\mathbf{X}^{(k)}_{\mathcal{O}}$ and the commutative relationship $\mathbf{C}^{(k)}\mathbf{S}^{*(k)} = \mathbf{S}^{*(k)}\mathbf{C}^{(k)}$ as described in Section II to recover the structure in $\mathbf{S}^{*(k)}_{\mathcal{O}}$. Given the problem setting, we can now formalize our joint topology inference problem in the presence of hidden nodes as follows.

Problem 1 *Given the sets $\{\mathbf{X}^{(k)}_{\mathcal{O}}\}_{k=1}^K$ of graph signal values at the observed nodes for each of the K graphs, recover $\{\mathbf{S}^{*(k)}_{\mathcal{O}}\}_{k=1}^K$ under the following assumptions:*
(AS1) the number of hidden nodes H is much smaller than the number of observed nodes, that is, $H \ll O$;
(AS2) the signals in $\mathbf{X}^{(k)}$ are realizations of a process that is stationary in $\mathbf{S}^{*(k)}$; and
(AS3) the GSOs $\mathbf{S}^{*(k)}$ and $\mathbf{S}^{*(k')}$ are sparse and have similar sparsity patterns.

We elaborate on the implications of the assumptions. The first assumption *(AS1)* ensures the tractability of the problem. When most of the nodes in the graph are observed, the covariance submatrix $\mathbf{C}^{(k)}_{\mathcal{O}}$ sufficiently characterizes the structure of $\mathbf{S}^{*(k)}_{\mathcal{O}}$. Importantly, under $H \ll O$, the matrix product $\mathbf{C}^{(k)}_{\mathcal{O}\mathcal{H}}\mathbf{S}^{*(k)}_{\mathcal{H}\mathcal{O}}$ is low-rank, a crucial result for inferring $\mathbf{S}^{*(k)}_{\mathcal{O}}$, which is also assumed in different single graph-learning approaches. Assumption *(AS2)* establishes a global relationship between the graph signals $\mathbf{X}^{(k)}$ and the unknown graph structure $\mathbf{S}^{*(k)}$, including both observed and hidden nodes. This assumption enables us to specify how the hidden nodes affect $\mathbf{X}^{(k)}$ by considering the connectivity between observed and hidden nodes encoded in $\mathbf{S}^{*(k)}_{\mathcal{O}\mathcal{H}}$ from (1) and the commutative relationship $\mathbf{C}^{(k)}\mathbf{S}^{*(k)} = \mathbf{S}^{*(k)}\mathbf{C}^{(k)}$. The final assumption *(AS3)* guarantees that all K graphs have similar edge connectivity patterns across all the shared node set $\mathcal{V}$. Not only can we then benefit from jointly inferring the observed subnetworks, but we may also share hidden node information across all K graphs during inference. We naturally expect that the support of $\mathbf{S}^{*(k)}_{\mathcal{O}}$ will be similar across all K graphs [22], [23], [35]; however, it is important to also exploit the edgewise similarity for $\mathbf{S}^{*(k)}_{\mathcal{O}\mathcal{H}}$ to account for connections between observed and hidden nodes.

Notice that for the simpler case where the set $\mathcal{H}$ of hidden nodes differs across graphs, *(AS3)* would allow us to exploit nodal observations from graph k that are hidden for graph k' to account for hidden nodes. However, in this work, we address the more challenging scenario in Problem 1, where there is a subset of nodes for which there are no direct observations for *any* graph. We rely on the statistical relationship between the graph signals and the graph topology to formulate a suitable optimization problem for jointly inferring the subnetworks in $\mathbf{S}^{*(k)}_{\mathcal{O}}$.

IV. Joint graph learning with latent variables as a convex optimization problem

Network topology inference with stationary graph signals commonly exploits the commutativity of the graph signal covariance matrices and the GSOs. We also adopt this approach; however, unlike previous works, we cannot directly apply the commutative relationship due to the presence of hidden nodes. We must revisit the commutativity of $\mathbf{C}^{(k)}$ and $\mathbf{S}^{*(k)}$ with the partitions in (1) before introducing our inference problem with stationary graph signals. From stationarity *(AS2)*, we know that $\mathbf{S}^{*(k)}\mathbf{C}^{(k)} = \mathbf{C}^{(k)}\mathbf{S}^{*(k)}$ for all $k = 1, \dots, K$. From (1) it then follows that

$$\mathbf{C}^{(k)}_{\mathcal{O}}\mathbf{S}^{*(k)}_{\mathcal{O}} - \mathbf{S}^{*(k)}_{\mathcal{O}}\mathbf{C}^{(k)}_{\mathcal{O}} = (\mathbf{P}^{*(k)})^\top - \mathbf{P}^{*(k)} \quad (2)$$

for all $k = 1, \dots, K$, where $\mathbf{P}^{*(k)} := \mathbf{C}_{\mathcal{O}\mathcal{H}}^{(k)} \mathbf{S}_{\mathcal{H}\mathcal{O}}^{*(k)}$. The right-hand side of (2) fully accounts for the influence of hidden nodes. When $\mathbf{P}^{*(k)}$ is known, estimating $\mathbf{S}_{\mathcal{O}}^{*(k)}$ relies solely on the commutator on the left-hand side. This is similar to traditional network inference with stationary graph signals, where we also know the value of the commutator $\mathbf{C}^{(k)}\mathbf{S}^{*(k)} - \mathbf{S}^{*(k)}\mathbf{C}^{(k)} = \mathbf{0}_{N\times N}$.

With the prior structural information in place, we can approach estimating the subnetworks from sample covariance submatrices $\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} = \frac{1}{R_k}\mathbf{X}_{\mathcal{O}}^{(k)}(\mathbf{X}_{\mathcal{O}}^{(k)})^\top$ by the following nonconvex optimization problem

$$\begin{aligned}\min_{\{\mathbf{S}_{\mathcal{O}}^{(k)},\mathbf{P}^{(k)}\}_{k=1}^K} \quad & \sum_{k=1}^K \alpha_k \|\mathbf{S}_{\mathcal{O}}^{(k)}\|_0 + \sum_{k<k'} \beta_{k,k'} \|\mathbf{S}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{(k')}\|_0 \\ & + \sum_{k=1}^K \gamma_k \|\mathbf{P}^{(k)}\|_{2,1} + \sum_{k<k'} \eta_{k,k'} \left\| \begin{bmatrix} \mathbf{P}^{(k)} \\ \mathbf{P}^{(k')} \end{bmatrix} \right\|_{2,1} \\ \text{s.t. } & \textstyle\sum_{k=1}^K \|\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\mathbf{S}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{(k)}\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} + \mathbf{P}^{(k)} - (\mathbf{P}^{(k)})^\top\|_F^2 \le \epsilon^2, \\ & \mathbf{S}_{\mathcal{O}}^{(k)} \in \mathcal{S}, \end{aligned} \tag{3}$$

where we have introduced auxiliary matrices $\{\mathbf{P}^{(k)}\}_{k=1}^K$ to account for the right hand side of (2). We first discuss (3) as it relates to $\{\mathbf{S}_{\mathcal{O}}^{(k)}\}_{k=1}^K$. The first two terms in the objective of (3) encourage sparse subnetworks with similar sparsity patterns as in *(AS3)*. The second constraint encourages valid GSOs for $\mathbf{S}_{\mathcal{O}}^{(k)}$. In this work, we let the GSOs denote adjacency matrices, so we define

$$\mathcal{S} := \left\{ \mathbf{S} \ : \ \mathbf{S} = \mathbf{S}^\top, \ \mathrm{diag}(\mathbf{S}) = \mathbf{0}, \ \textstyle\sum_j \mathbf{S}_{j1} = 1 \right\}, \tag{4}$$

where $\{\mathbf{S}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ denote valid submatrices of nontrivial adjacency matrices, that is, $\mathbf{S}_{\mathcal{O}}^{(k)} \neq \mathbf{0}_{O\times O}$. While we select adjacency matrices as GSOs, problem (3) accommodates other GSOs, such as the graph Laplacian [16], under minor modifications.

We next discuss the auxiliary matrices $\{\mathbf{P}^{(k)}\}_{k=1}^K$. The first constraint encourages the commutativity in (2) with $\mathbf{P}^{(k)}$ as an approximation of $\mathbf{P}^{*(k)} = \mathbf{C}_{\mathcal{O}\mathcal{H}}^{(k)}\mathbf{S}_{\mathcal{H}\mathcal{O}}^{*(k)}$ to avoid a bilinear formulation. As will be discussed in Section V, the upper bound ϵ accounts for both the sample covariance submatrix error and the difference between $\mathbf{P}^{(k)}$ and $\mathbf{P}^{*(k)}$. Thus, similarly to [35], we introduce the low-rank matrices $\mathbf{P}^{(k)}$ to replace entities that depend on hidden nodes. However, instead of using the standard convex surrogate for low-rankness given by the nuclear norm, we rely on the $\ell_{2,1}$ to impose additional structure on $\mathbf{P}^{(k)}$ based on the assumptions in Problem 1.

Precisely, the last two terms in the objective apply a group Lasso penalty via the $\ell_{2,1}$ norm [34], which evaluates the ℓ_1 norm of the vector containing the ℓ_2 norm of each column of the input matrix, that is, $\|\mathbf{P}^{(k)}\|_{2,1} = \sum_{i=1}^O \|\mathbf{P}_{\cdot,i}^{(k)}\|_2$. Recall that since $H \ll O$ by *(AS1)*, the matrix $\mathbf{P}^{*(k)}$ is not only low-rank but has sparse columns, hence the third term in the objective applying the $\ell_{2,1}$ norm to encourage column-sparsity in $\mathbf{P}^{(k)}$. While low-rank constraints are commonly implemented with the convex nuclear norm penalty [32], where solutions with sparse singular values are sought, we simultaneously promote low-rankness while encouraging column sparsity by the group Lasso penalty. Additionally, since the networks are assumed to have similar sparsity patterns by *(AS3)*, we expect that the column sparsity patterns of $\mathbf{P}^{*(k)}$ across networks will be similar, hence the fourth term in the objective.

As is common with optimization problems for sparse network inference, we introduce a convex relaxation of (3) that enjoys efficient solvability and theoretical guarantees. Our convex formulation is

$$\begin{aligned}\min_{\{\mathbf{S}_{\mathcal{O}}^{(k)},\mathbf{P}^{(k)}\}_{k=1}^K} \quad & \sum_{k=1}^K \alpha_k \|\mathbf{S}_{\mathcal{O}}^{(k)}\|_1 + \sum_{k<k'} \beta_{k,k'} \|\mathbf{S}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{(k')}\|_1 \\ & + \sum_{k=1}^K \gamma_k \|\mathbf{P}^{(k)}\|_{2,1} + \sum_{k<k'} \eta_{k,k'} \left\| \begin{bmatrix} \mathbf{P}^{(k)} \\ \mathbf{P}^{(k')} \end{bmatrix} \right\|_{2,1} \\ \text{s.t. } & \textstyle\sum_{k=1}^K \|\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\mathbf{S}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{(k)}\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} + \mathbf{P}^{(k)} - (\mathbf{P}^{(k)})^\top\|_F^2 \le \epsilon^2, \\ & \mathbf{S}_{\mathcal{O}}^{(k)} = (\mathbf{S}_{\mathcal{O}}^{(k)})^\top, \ \mathrm{diag}(\mathbf{S}_{\mathcal{O}}^{(k)}) = \mathbf{0}, \ \forall k = 1, \dots, K, \\ & \textstyle\sum_j [\mathbf{S}_{\mathcal{O}}^{(1)}]_{j1} = 1, \end{aligned} \tag{5}$$

where we have removed the nonconvexities in (3) by substituting the ℓ_0 norms in the objective with convex ℓ_1 norms. We further specified the constraints according to (4) for valid adjacency submatrices. While the last constraint is valid to preclude trivial adjacency submatrices, it would not be viable for graph Laplacians as GSOs. However, the theoretical results in Section V still hold for graph Laplacian GSOs by replacing the last constraint in (4) to enforce valid graph Laplacian submatrices.

V. Theoretical results

We formalize the viability of the convex relaxation in (5) by presenting conditions under which the solutions to (3) and (5) are equivalent. We also compute an upper bound on the error of the solution to (5) and apply the bound to evaluate the effectiveness of (5) at accounting for hidden nodes.

A. Sparsity of the convex relaxation

We first introduce the following definitions to rewrite the optimization problems in (3) and (5) in vector form. Let the vectors $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{\beta} \in \mathbb{R}^{K(K-1)/2}$ collect values of α_k and $\beta_{k,k'}$, respectively. Let $\mathcal{L}' := \mathcal{L}^{(1)} \cup \cdots \cup \mathcal{L}^{(K)}$, where $\mathcal{L}^{(k)} := \{i = j + (k-1)O^2 : j \in \mathcal{L}\}$ for $\mathcal{L}$ containing indices for a O^2-length vector (corresponding to the vector form of an $O\times O$ matrix) as described in Section II. We define the directed difference matrix $\mathbf{Z} := [\mathbf{1}_K^\top \otimes -\mathbf{I}_K]_{\cdot,\mathcal{L}} + [\mathbf{I}_K \otimes \mathbf{1}_K^\top]_{\cdot,\mathcal{L}}$, where $\mathcal{L}$ contains indices for a K^2-length vector. We can then introduce the matrix $\boldsymbol{\Psi} := 2[\boldsymbol{\Psi}_0]_{\cdot,\mathcal{L}'}$ associated with the objectives of (3) and (5), where

$$\boldsymbol{\Psi}_0 := \begin{bmatrix} \mathrm{diag}(\boldsymbol{\alpha}) \otimes \mathbf{I}_{O^2} \\ \mathrm{diag}(\boldsymbol{\beta})\mathbf{Z}^\top \otimes \mathbf{I}_{O^2} \end{bmatrix}.$$

For the first constraint of (3) and (5), we introduce $\boldsymbol{\Sigma} := \mathrm{blockdiag}(\boldsymbol{\Sigma}^{(1)}, \dots, \boldsymbol{\Sigma}^{(K)})$, where $\boldsymbol{\Sigma}^{(k)} := [\boldsymbol{\Sigma}_0^{(k)}]_{\cdot,\mathcal{L}} + [\boldsymbol{\Sigma}_0^{(k)}]_{\cdot,\mathcal{U}}$ and $\boldsymbol{\Sigma}_0^{(k)} = (-\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} \oplus \hat{\mathbf{C}}_{\mathcal{O}}^{(k)})$ for all $k = 1, \dots, K$, and $\mathcal{L}$ and $\mathcal{U}$ for $\boldsymbol{\Sigma}^{(k)}$ return entries of a vector of length O^2. Furthermore, let $\mathbf{Q}$ be a commutation matrix such that for any square matrix $\mathbf{Y}$, we have that $\mathrm{vec}(\mathbf{Y}^\top) = \mathbf{Q}\mathrm{vec}(\mathbf{Y})$, and let

$\mathbf{M} = \text{blockdiag}(\mathbf{I}_{O^2} - \mathbf{Q}, \ldots, \mathbf{I}_{O^2} - \mathbf{Q})$ with K diagonal blocks. Let $\mathcal{E}^{(k,i)} = \{(k-1)O^2 + (i-1)O + j\}_{j=1}^{O}$ be index sets for all $k = 1, \ldots, K$ and $i = 1, \ldots, O$. Based on this, define $\mathcal{E}^{(k,k',i)} = \mathcal{E}^{(k,i)} \cup \mathcal{E}^{(k',i)}$ for every $k, k' = 1, \ldots, K$ with $k < k'$, where $\mathcal{E}^{(k,i)}$ corresponds to the indices of the i-th column in the vectorized version of the matrix $\mathbf{P}^{(k)}$ and $\mathcal{E}^{(k,k',i)}$ to the indices of the i-th columns of the vectorized versions of $\mathbf{P}^{(k)}$ and $\mathbf{P}^{(k')}$.

With the following vectorizations,

$$\mathbf{s} = [\text{vec}(\mathbf{S}_{\mathcal{O}}^{(1)})_{\mathcal{L}}^{\top}, \ \cdots, \ \text{vec}(\mathbf{S}_{\mathcal{O}}^{(K)})_{\mathcal{L}}^{\top}]^{\top} \in \mathbb{R}^{KO(O-1)/2}, \quad (6)$$

$$\mathbf{p} = [\text{vec}(\mathbf{P}^{(1)})^{\top}, \ \cdots, \ \text{vec}(\mathbf{P}^{(K)})^{\top}]^{\top} \in \mathbb{R}^{KO^2}, \quad (7)$$

we may rewrite the optimization problem (3) as

$$\begin{aligned}\{\mathbf{s}', \mathbf{p}'\} = \underset{\{\mathbf{s},\mathbf{p}\}}{\text{argmin}} \quad & \|\mathbf{\Psi s}\|_0 + \sum_{k=1}^{K}\sum_{i=1}^{O} \gamma_k \|\mathbf{p}_{\mathcal{E}^{(k,i)}}\|_2 \\ & + \sum_{k<k'}\sum_{i=1}^{O} \eta_{k,k'} \|\mathbf{p}_{\mathcal{E}^{(k,k',i)}}\|_2 \\ \text{s.t.} \quad & \|\mathbf{\Sigma s} + \mathbf{M p}\|_2 \leq \epsilon, \ (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^{\top}\mathbf{s} = 1 \end{aligned} \quad (3')$$

and (5) as

$$\begin{aligned}\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\} = \underset{\{\mathbf{s},\mathbf{p}\}}{\text{argmin}} \quad & \|\mathbf{\Psi s}\|_1 + \sum_{k=1}^{K}\sum_{i=1}^{O} \gamma_k \|\mathbf{p}_{\mathcal{E}^{(k,i)}}\|_2 \\ & + \sum_{k<k'}\sum_{i=1}^{O} \eta_{k,k'} \|\mathbf{p}_{\mathcal{E}^{(k,k',i)}}\|_2 \\ \text{s.t.} \quad & \|\mathbf{\Sigma s} + \mathbf{M p}\|_2 \leq \epsilon, \ (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^{\top}\mathbf{s} = 1. \end{aligned} \quad (5')$$

We further denote $\mathcal{J}$ as $\text{supp}(\mathbf{\Psi s}')$ and $\mathcal{I}$ as $\text{supp}(\mathbf{s}')$, where $\text{supp}(\mathbf{y})$ denotes the support of the vector $\mathbf{y}$. With the above definitions in place, we have the following result.

Theorem 1. *Assume that problem* (5') *is feasible. The solution* $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$ *of* (5') *is equivalent to the solution* $\{\mathbf{s}', \mathbf{p}'\}$ *of* (3') *if the following two conditions are satisfied:*

1) $\mathbf{\Sigma}_{\cdot,\mathcal{I}}$ *is full column rank; and*
2) *There exist constants* $\psi, C_s > 0$ *such that*

$$\|\mathbf{\Psi}_{\mathcal{J}^c,\cdot}(\mathbf{T}_1 - \mathbf{T}_2)\mathbf{\Psi}_{\mathcal{J},\cdot}^{\top}\|_{\infty} < 1,$$

where

$$\begin{aligned}\mathbf{T}_1 &:= \big(\psi^{-2}(\mathbf{\Sigma}^{\top}\mathbf{\Sigma} + 2\epsilon^2 C_s^{-2}\mathbf{I}_{KO(O-1)/2}) \\ &\qquad + \mathbf{\Psi}_{\mathcal{J}^c,\cdot}^{\top}\mathbf{\Psi}_{\mathcal{J}^c,\cdot}\big)^{-1}, \\ \mathbf{T}_2 &:= \frac{\mathbf{T}_1(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^{\top}\mathbf{T}_1}{(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^{\top}\mathbf{T}_1(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})}. \end{aligned}$$

The proof of Theorem 1 can be found in Appendix A, but we also provide a summary here. To decouple the joint optimization of $\mathbf{s}$ and $\mathbf{p}$, we consider an alternating minimization algorithm, permitting separate analysis of $\mathbf{s}$-subproblems and $\mathbf{p}$-subproblems at each iteration. Proximal alternating minimization [46], an iterative optimization algorithm, applied to (3') and (5') can be shown to converge to the original solutions $\{\mathbf{s}', \mathbf{p}'\}$ and $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$, respectively. We then can show that the $\mathbf{p}$-subproblems for (3') and (5') are equivalent for every iteration, and therefore $\mathbf{p}' = \hat{\mathbf{p}}$. When the iterations grow sufficiently large for convergence, the $\mathbf{s}$-subproblems for (3') and (5') are equivalent under the conditions of Theorem 1, so $\mathbf{s}' = \hat{\mathbf{s}}$.

Under the sufficient conditions of Theorem 1, the convex relaxation in (5) enjoys recovery of the sparsest solution of (3) even in the presence of hidden nodes. Note that this result differs significantly from that of Theorem 1 in [23] due to the presence of another variable $\mathbf{p}$ that is not associated with an entrywise sparsity penalty. Condition *1)* of Theorem 1 guarantees that the solution to (5) is unique, and condition *2)* permits the existence of a dual certificate that ensures that the solutions to (5) and (3) are equivalent [23], [47]. Thus, under the conditions of Theorem 1, the ℓ_1 norm does not introduce any estimation error for obtaining the sparsest GSO submatrix estimates, and we need only consider the distortion from the sample covariance submatrices $\{\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\}_{k=1}^{K}$ and auxiliary matrices $\{\hat{\mathbf{P}}^{(k)}\}_{k=1}^{K}$ obtained from (5).

B. Robust recovery under hidden nodes

By Theorem 1, we can guarantee under mild conditions when the solution to (5) is equivalent to the sparsest solution from (3). Therefore, to evaluate the efficacy of our method in estimating the true GSO submatrices $\{\mathbf{S}_{\mathcal{O}}^{*(k)}\}_{k=1}^{K}$, we need only consider the estimation error of (5). In the sequel, we derive an upper bound on the distortion between the true GSO submatrices $\{\mathbf{S}_{\mathcal{O}}^{*(k)}\}_{k=1}^{K}$ and the estimated ones $\{\hat{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^{K}$ obtained from (5). Let $\mathbf{s}^*$ be the vectorization of the true GSO submatrices $\{\mathbf{S}_{\mathcal{O}}^{*(k)}\}_{k=1}^{K}$ as in (6). We define $\mathcal{K}$ as $\text{supp}(\mathbf{\Psi s}^*)$, and we let $R := \sum_{k=1}^{K} R_k$ and $\omega := \max_{k=1,\ldots,K} \omega_k$, where $\omega_k := \max\{\max_i [\mathbf{C}_{\mathcal{O}}^{(k)}]_{ii}, \max_i [\mathbf{S}_{\mathcal{O}}^{*(k)}\mathbf{C}_{\mathcal{O}}^{(k)}\mathbf{S}_{\mathcal{O}}^{*(k)}]_{ii}\}$. We present our main result on the performance of our proposed method.

Theorem 2. *Let* $\{\hat{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^{K}$ *be the estimated subnetworks obtained from* (5) *with* $\epsilon = \epsilon_R + \alpha$ *for*

$$\alpha^2 = \sum_{k=1}^{K} \left\| \big(\hat{\mathbf{P}}^{(k)} - (\hat{\mathbf{P}}^{(k)})^{\top}\big) - \big(\mathbf{P}^{*(k)} - (\mathbf{P}^{*(k)})^{\top}\big) \right\|_F^2$$

and $\epsilon_R \geq C_1 O\omega\sqrt{(K \log O)/R}$ *for some constant* $C_1 > 0$. *Under the following four conditions,*

1) $K = o(\log O)$;
2) $R_1 \asymp R_2 \asymp \cdots \asymp R_K$;
3) $\log O = o(\min\{R/(K^7(\log R)^2), (R/K^7)^{1/3}\})$; *and*
4) $\mathbf{\Sigma}$ *is full column rank;*

with probability at least $1 - e^{-C_2 \log O}$ *for some constant* C_2 *we have that*

$$\begin{aligned}&\sum_{k=1}^{K} \|\hat{\mathbf{S}}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}\|_1 \leq \tau(\epsilon_R + \alpha), \\ &\textit{where } \tau = \frac{4\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})\|\mathbf{\Psi}^{\dagger}\|_1}{\sigma_{\min}(\mathbf{\Sigma})}(2 + \sqrt{|\mathcal{K}|}). \end{aligned} \quad (8)$$

The proof of Theorem 2 can be found in Appendix B. In brief, we first apply the commutative relationship described

in Section II to show that $\{\mathbf{s}^*, \hat{\mathbf{p}}\}$ is a feasible solution to (5'). We can then bound the ℓ_1-norm difference between the vectorization of the true GSOs $\mathbf{s}^*$ and the estimated one $\hat{\mathbf{s}}$ based on the commutativity constraint, $\epsilon = \epsilon_R + \alpha$.

Theorem 2 presents an upper bound on the estimation error of (5). If K and O are fixed, then as the number of observed graph signals R increases, the sample covariance submatrices $\{\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ approach the true covariance submatrices, and the first term $\tau\epsilon_R$ in the upper bound in (8) becomes negligible. With enough observed graph signals, the error primarily depends on the second term $\tau\alpha$, which denotes the approximation error of $\{\hat{\mathbf{P}}^{(k)}\}_{k=1}^K$, the crux of our proposed method. If (5) is effective at enforcing $\mathbf{P}^{(k)}$ to share structural characteristics of $\mathbf{C}_{\mathcal{O}\mathcal{H}}^{(k)}\mathbf{S}_{\mathcal{H}\mathcal{O}}^{*(k)}$ such that they are close, then the estimation of the GSO submatrices $\mathbf{S}_{\mathcal{O}}^{*(k)}$ becomes easier according to (8). Furthermore, as $\mathbf{P}^{(k)}$ becomes a more accurate approximation of $\mathbf{P}^{*(k)}$, the estimation accuracy of $\hat{\mathbf{S}}_{\mathcal{O}}^{(k)}$ improves increasingly when compared to estimating $\mathbf{S}_{\mathcal{O}}^{*(k)}$ while ignoring the presence of hidden nodes. We formalize this statement in the following result that characterizes the effectiveness of our proposed formulation with respect to the auxiliary matrices $\{\mathbf{P}^{(k)}\}_{k=1}^K$.

Corollary 1. *Let the naive subnetwork estimates considering only observed nodes be denoted as $\{\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ [23], which we define as the solution to* (5) *while fixing $\mathbf{P}^{(k)} = \mathbf{0}_{O\times O}$ for every $k = 1, 2, \ldots, K$, and we let $\epsilon = \epsilon_R$, where $\epsilon_R \geq C_1 O\omega\sqrt{(K\log O)/R}$ for some constant $C_1 > 0$, and $\gamma_k = 0$, $\eta_{k,k'} = 0$ for every $k, k' = 1, 2, \ldots, K$ and $k < k'$. Additionally, let $\tilde{\mathbf{s}}$ be the vectorization as in* (6) *of $\{\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ and define δ as*

$$\delta^2 = \sum_{k=1}^K \|\mathbf{P}^{*(k)} - (\mathbf{P}^{*(k)})^\top\|_F^2.$$

Then, we have that

$$\sum_{k=1}^K \|\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}\|_1 \leq (\tau + \tau')(\epsilon_R + \tfrac{1}{2}\delta),$$
$$\text{where } \tau = \frac{4\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})\|\mathbf{\Psi}^\dagger\|_1}{\sigma_{\min}(\mathbf{\Sigma})}(2 + \sqrt{|\mathcal{K}|})$$
$$\text{and } \tau' = \frac{2\rho KO(O-1)(1+\sqrt{|\mathcal{K}|})\sigma_{\max}(\mathbf{\Psi})\|\mathbf{\Psi}^\dagger\|_1}{\sigma_{\min}(\mathbf{\Sigma})} \quad (9)$$

for some $\rho \in [0, 1]$. Furthermore, we have that if

$$\sum_{k=1}^K \left\| \left(\hat{\mathbf{P}}^{(k)} - (\hat{\mathbf{P}}^{(k)})^\top\right) - \left(\mathbf{P}^{*(k)} - (\mathbf{P}^{*(k)})^\top\right) \right\|_F^2 \leq \left(\frac{\tau'}{\tau}\right)^2 \epsilon_R^2 + \left(\frac{\tau+\tau'}{2\tau}\right)^2 \sum_{k=1}^K \left\| \mathbf{P}^{*(k)} - (\mathbf{P}^{*(k)})^\top \right\|_F^2, \quad (10)$$

then the error bound in (8) *is lower than the error bound in* (9).

The proof of Corollary 1 can be found in Appendix C, which follows a similar procedure to the proof of Theorem 2. Corollary 1 demonstrates the criticality of accounting for hidden nodes. We describe these implications more intuitively here. First, as discussed following Theorem 2, we note that as $\hat{\mathbf{P}}^{(k)}$ approximates $\mathbf{P}^{*(k)}$ more accurately, we achieve greater improvement over $\{\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ from our proposed inference problem (5). Indeed, as the matrix difference $(\hat{\mathbf{P}}^{(k)})^\top - \hat{\mathbf{P}}^{(k)}$ approaches the right-hand side of (2), we remove the influence of the hidden nodes on the estimation of the observed submatrices. Second, note that the second term in the upper bound of (10) is proportional to δ, which measures the influence of the hidden nodes on the observed nodes in the stationary graph signal regime. When δ is negligible, the hidden nodes have little effect on the observed nodes, and the inclusion of $\{\mathbf{P}^{(k)}\}_{k=1}^K$ in the inference process may affect performance detrimentally. However, as δ increases, the need to account for the right-hand side of (2) becomes crucial. We verify this comparison of (5) and the naive solution $\{\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ with synthetic simulations in Section VI.

VI. Numerical evaluation

We introduce several experiments to assess the performance of the proposed network topology inference method. The experiments employ synthetic and real-world data and compare the quality of the graphs estimated by different algorithms. For the k-th graph, we compute the normalized error between the true $\mathbf{S}_{\mathcal{O}}^{*(k)}$ and the estimated $\hat{\mathbf{S}}_{\mathcal{O}}^{(k)}$ as

$$\text{nerr}(\mathbf{S}_{\mathcal{O}}^{*(k)}, \hat{\mathbf{S}}_{\mathcal{O}}^{(k)}) = \frac{\|\mathbf{S}_{\mathcal{O}}^{*(k)} - \hat{\mathbf{S}}_{\mathcal{O}}^{(k)}\|_F^2}{\|\mathbf{S}_{\mathcal{O}}^{*(k)}\|_F^2}, \quad (11)$$

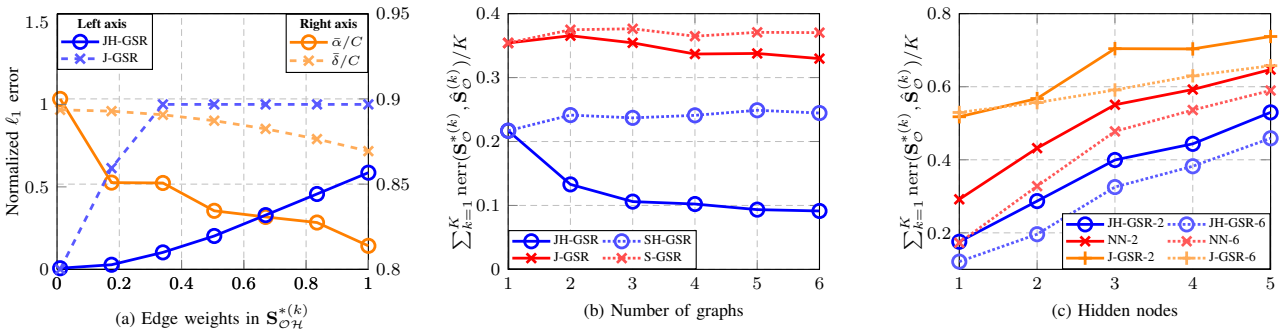
and then report the average across the K graphs being estimated, i.e., $\frac{1}{K}\sum_{k=1}^K \text{nerr}(\mathbf{S}_{\mathcal{O}}^{*(k)}, \hat{\mathbf{S}}_{\mathcal{O}}^{(k)})$. The code for the proposed method and the experiments is available on GitHub[1].

A. Synthetic experiments

We rely on synthetic graphs and signals to assess how different elements impact the performance of the proposed approach. Unless specified otherwise, in the following experiments we consider $K = 3$ graphs with $N = 20$ nodes from which $O = 19$ are observed. The graph $\mathcal{G}^{(1)}$ is sampled from an Erdős-Rényi (ER) random graph model with a link probability of $p = 0.2$, and the related graphs are created by randomly rewiring a fixed number of edges. Stationary graph signals are generated by diffusing a white input signal across the graph, i.e., $\mathbf{x} = \mathbf{H}\mathbf{w}$, where the coefficients of $\mathbf{H}$ are drawn from a uniform distribution and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Varying the effect of hidden nodes. We start by illustrating the result in (10) that expresses when it is beneficial to incorporate $\mathbf{P}^{(k)}$ for hidden nodes. To this end, we estimate $K = 3$ networks from perfectly known covariance submatrices $\mathbf{C}_{\mathcal{O}}^{(k)}$ so $\epsilon_R = 0$ [cf. (10)], to assess only the effects of $\mathbf{P}^{(k)}$ and the hidden nodes $\mathcal{H}$, characterized respectively by α from Theorem 2 and δ from Corollary 1. We compare two network inference methods: (i) "JH-GSR", which denotes the method in (5) that accounts for hidden nodes, and (ii) "J-GSR", which denotes the method described in Corollary 1 that ignores hidden variables [23]. Fig. 1a shows the network estimation

[1] https://github.com/reysam93/hidden_joint_inference



(a) Edge weights in $\mathbf{S}_{\mathcal{OH}}^{*(k)}$ (b) Number of graphs (c) Hidden nodes

Fig. 1: We test the performance of the proposed network topology inference in different settings. (a) Evaluation of the performance of graph inference accounting for hidden nodes via (5) and graph inference ignoring hidden nodes as described in Corollary 1 as the weights of edges between observed and hidden nodes increase. (b) Evaluation of the influence of increasing the number of graphs being estimated. (c) Evaluation of the detrimental effects of increasing the number of hidden nodes. The experiments consider different graph learning alternatives and the reported results are the average error of 100 independent realizations.

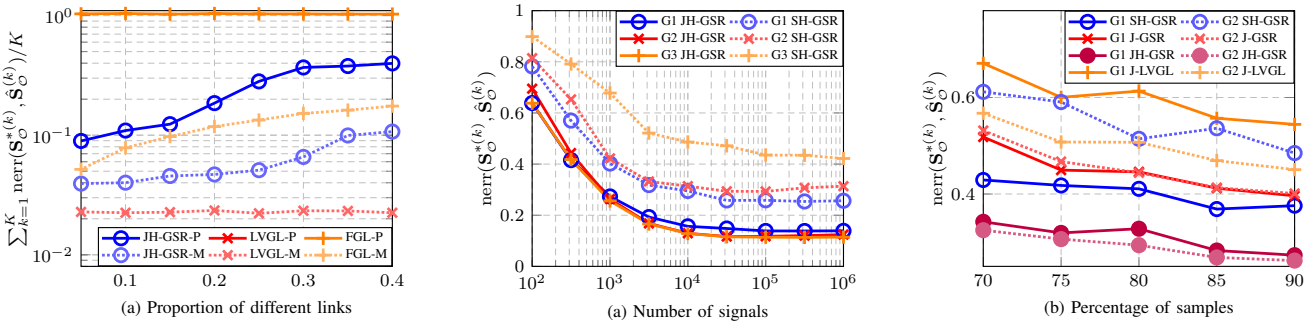
error as the edge weights connecting observed nodes and hidden nodes increase, that is, as nonzero entries in $\mathbf{S}_{\mathcal{OH}}^{*(k)}$ grow larger. While the GSO sparsity patterns do not change, the hidden node influence δ increases with the edge weights in $\mathbf{S}_{\mathcal{OH}}^{*(k)}$. To measure performance that is consistent with Corollary 1, we report the average error across all K graphs as the normalized ℓ_1-norm difference, equivalent to computing (11) with the ℓ_1 norm replacing the squared Frobenius norm. We let $\epsilon = 10^{-8}$ for the first constraint in (5); however, the solution to the naive problem with $\mathbf{P}^{(k)} = \mathbf{0}_{O\times O}$ may not be feasible. Indeed, when ϵ is small enough, it may be impossible to obtain a feasible solution $\{\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^{K}$ such that all constraints hold. In such a case where the solution is infeasible, we let its error be 1. Along with network estimation error, we compare in Fig. 1a normalized values of α and δ to evaluate when the result in (10) holds. In particular, we let $\bar{\alpha} := \sum_k \operatorname{nerr}(\mathbf{P}^{*(k)}, (\mathbf{P}^{*(k)})^\top + \hat{\mathbf{P}}^{(k)} - (\hat{\mathbf{P}}^{(k)})^\top)/K$ and $\bar{\delta} := \sum_k \operatorname{nerr}(\mathbf{P}^{*(k)}, (\mathbf{P}^{*(k)})^\top)/K$. Since we need only consider which value is greater, we plot $\bar{\alpha}/C$ and $\bar{\delta}/C$ for some constant $C > 0$ such that the values are between 0 and 1.

When the edge weight is 0, the hidden nodes are decoupled from the network and thus have no effect on the observed nodes, and indeed "J-GSR" perfectly recovers the true networks. For zero-valued edge weights in $\mathbf{S}_{\mathcal{OH}}^{*(k)}$, we observe $\alpha \geq \delta$, where "JH-GSR" is comparable but not superior to "J-GSR". As the edge weight increases and becomes nonnegligible, the effect of the hidden nodes increases, and we observe in Fig. 1a that $\alpha < \delta$ for all nonzero edge weights and "JH-GSR" consistently outperforms "JH-GSR" as expected from (10). We thus validate the necessity of our proposed method, where as the influence of hidden nodes increases, we must account for their presence to maintain a satisfactory estimation error.

Varying the number of graphs. We next assess the benefits of considering a joint network topology inference approach when several graphs need to be learned. To that end, Fig. 1b illustrates the normalized error computed according to (11) as the number of graphs K being estimated increases. The performance of "JH-GSR" is compared with (i) "S-GSR", the network topology inference method from stationary observations [16] where graphs are learned individually and the presence of hidden variables is ignored; "SH-GSR", a generalization of (i) that takes into account the influence of hidden variables [32]; and (iii) "J-GSR" as in Fig. 1a. Looking at the results, we observe that "JH-GSR" outperforms the alternatives, showcasing the benefits of harnessing the graph similarity while accounting for the influence of the hidden nodes. We also observed that the joint approaches achieve a lower error when more than one graph is being estimated, and furthermore, that the benefits of the joint approaches increase with K. Lastly, Fig. 1b also shows that for the setup at hand, ignoring the influence of hidden nodes results in a worse performance than ignoring the relation across networks, which is studied in more detail in the following experiment.

Varying the number of hidden nodes. The results in Fig. 1c investigate the detrimental influence of the presence of hidden nodes in the network topology inference task. We examine fixed-size graphs with $N = 20$ nodes and increase the number of hidden nodes H as shown in the x-axis. We evaluate the performance of (i) our proposed method, "JH-GSR", (ii) an alternative implementation of our method replacing the group Lasso penalty by the nuclear norm, "NN", and (iii) the joint network topology inference ignoring the presence of hidden nodes, "J-GSR" [23]. Then, for each baseline, we consider the estimation of either 2 or 6 graphs. First, from Fig. 1c, it can be seen that increasing the number of hidden nodes renders the inference problem more challenging and, moreover, that ignoring the presence of hidden nodes results in poor performance. Second, the superior performance of "JH-GSR" over "NN" supports our initial intuition that the group Lasso penalty is better suited to capture the structure of the problem at hand. Furthermore, we also observe that estimating 6 graphs leads to a better performance than estimating 2, a behavior aligned with the previous experiment.

Varying graph similarity. The last experiment involving synthetic data tests the impact of *(AS3)*, a critical assumption in joint graph learning. More precisely, we consider estimating $K = 3$ graphs as the proportion of different edges increases, i.e., as the graphs become more dissimilar. The errors of the estimated graphs are depicted in Fig. 2a, where we compare the



(a) Proportion of different links (a) Number of signals (b) Percentage of samples

Fig. 2: We test the performance of the proposed network topology inference in settings with synthetic and real-world data. (a) Evaluation of the impact of the graph similarity in joint network topology inference methods. This experiment considers different graph learning alternatives and the reported results are the average error of 100 independent realizations. (b) Error estimating three graphs considering either a joint or a separate method. Graphs are obtained from the students of the University of Ljubljana dataset. (c) Error estimating two graphs from voting signals considering different approaches.

performance of "JH-GSR" with (i) "LVGL", a graphical Lasso algorithm modeling the presence of hidden nodes [27]; and (ii) "FGL", a joint graphical Lasso algorithm [22]. Moreover, since graphical Lasso algorithms assume that the observations are drawn from a GMRF, we consider two different types of signals. Signals sampled from a GMRF are denoted as "M", and signals generated as the diffusion of a white input via a polynomial of the GSO are denoted as "P". As expected from *(AS3)*, Fig. 2a shows that the performance of joint methods, "JH-GSR" and "FGL", deteriorates as we consider a higher number of different links. For the two signal models, we observe that "JH-GSR-M" is superior to "JH-GSR-P" since the GMRF model is a simpler special case of graph stationarity that is less sensitive to hidden nodes. Interestingly, "JH-GSR-M" also outperforms "FGL-M", although the latter is a method tailored for GMRF observations, showcasing the more general nature of the stationary model and the importance of accounting for the presence of hidden nodes. In contrast, we observe that graphical models are incapable of estimating graphs from stationary observations, and we note that "LVGL-P" is not included in the figure due to its high error.

B. Application to real-world graphs

In addition to the synthetic data where we know the model relating the networks and the observed graph signals, we assess our proposed method with real-world data to demonstrate its efficacy in several scenarios, including those where the stationarity assumption is not explicitly enforced.

Students dataset. The following experiment combines real-world graphs with synthetic signals. This mixed approach allows us to investigate the applicability of the proposed method to real-world graphs while ensuring that the observed signals are stationary. We employed three graphs defined on a common set of 32 nodes, where nodes represent students from the University of Ljubljana, and the different graphs encode various types of interactions among the students[2]. The results are displayed in Fig. 2b, where we observe the error of the recovered graphs as the number of samples increases. The error reported is the average of 50 realizations of random stationary graph signals, with only one hidden node considered. For each of the three graphs, we evaluate the performance of both the joint and the separate estimation methods, "JH-GSR" and "SH-GSR". From the results, it is evident that the recovery of all three graphs significantly improves with a joint approach, demonstrating the benefits of leveraging the existing relationship between the networks.

Learning multiple observed graphs from voting data. Finally, we close with an experiment aimed at learning two related political graphs from voting data[3]. More specifically, we consider 25 cantons of Switzerland as the nodes of the graph and the percentage of votes in favor of 185 initiatives submitted between 2000 and 2020 as the signals. Our goal then is to infer the political graph of Switzerland for two consecutive periods of time. Intuitively, although political representation may evolve with time, this process is typically slow and, hence, the two graphs are expected to be closely related. We validate the estimations via ground truth graphs whose links reflect the political preferences of the cantons, which are obtained by performing separate inference of both graphs with all available signals. We consider $H = 2$ hidden nodes and estimate the $K = 2$ graphs varying the percentage of available signals from 70% to 90%. We compare the proposed algorithm, "JH-GSR", with three alternative methods: "J-GSR", "SH-GSR", and "J-LVGL" from [35].

The estimation error of the two graphs using the four methods is shown in Fig. 2c. Since the number of available signals for the second graph is considerably smaller than the signals available for the first graph, we observe a much larger estimation error for the second graph when the separate approach "SH-GSR" is employed. In contrast, for the joint estimation method "J-GSR", we observe that errors are similar for both graphs and inferior on average compared to "SH-GSR". This behavior illustrates that harnessing the similarity of the graphs results in an improvement in performance since it allows sharing common learned structures across graphs. Moreover, we observe that "JH-GSR" outperforms both "SH-GSR" and "J-GSR" since, in addition to being a joint ap-

[2]Original data available at http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:pajek:students

[3]Original data available at https://swissvotes.ch/page/home

proach, it takes into account the influence of the hidden nodes. We also compare "JH-GSR" with "J-LVGL", both of which perform joint network inference while accounting for hidden nodes. However, we find that "JH-GSR" is drastically superior due to complexities in the data structure that "J-LVGL" cannot capture accurately. Indeed, the stationary model subsumes the GMRF model while allowing for more complex statistical relationships between the graph topology and the signals.

To summarize, it is not only crucial to account for the presence of hidden nodes but, when several related graphs are involved, it is also important to exploit the similarity between both observed and hidden nodes. This becomes particularly relevant when data is limited to a subset of the graphs, as demonstrated in the improved estimation of the second graph when considering joint network inference methods.

VII. Conclusion

In this paper, we presented a method to infer multiple networks on the same node set in the presence of hidden nodes. To characterize the effect of the hidden nodes, we assumed that graph signals were stationary on their respective networks. By the inherent block structure of the covariance matrix $\mathbf{C}^{(k)}$ and the GSO $\mathbf{S}^{*(k)}$ of the k-th network, we introduced a set of auxiliary matrices $\mathbf{P}^{(k)}$ to account for the effect of hidden nodes in the relationship $\mathbf{C}^{(k)}\mathbf{S}^{*(k)} = \mathbf{S}^{*(k)}\mathbf{C}^{(k)}$ stemming from the stationarity assumption. By prior assumptions on structure and stationarity, we derive characteristics of $\mathbf{P}^{(k)}$ that permit us to form an optimization problem that performs network inference while accounting for the presence of hidden nodes. Moreover, we verified that the estimation of the sparsest networks is equivalent to a computationally feasible convex relaxation under mild conditions. We further demonstrated a bound on the error of our proposed method dependent on the error due to the sample covariance matrices and $\mathbf{P}^{(k)}$. The performance of our method was evaluated in multiple synthetic and real-world datasets in comparison with other baseline methods, and we also verified the improvement in estimation due to the incorporation of $\mathbf{P}^{(k)}$.

Appendix A
Proof of Theorem 1

We first combine the last two terms in the objective functions of (3') and (5') by defining the combined index set $\mathcal{E} := \bigcup_{i=1}^{O}\{\mathcal{E}^{(k,i)}\}_{k=1}^{K} \cup \{\mathcal{E}^{(k,k',i)}\}_{k<k'}$ and parameters $\{\eta'_g\}_{g\in\mathcal{E}}$ such that $\eta'_{\mathcal{E}^{(k,i)}} = \gamma_k$ and $\eta'_{\mathcal{E}^{(k,k',i)}} = \eta_{k,k'}$ for every $k, k' = 1, \dots, K$ such that $k < k'$ and $i = 1, \dots, O$.

Let us consider solving (3') by proximal alternating minimization [46] with

$$\begin{aligned}\mathbf{p}'^{(t)} &= \operatorname*{argmin}_{\mathbf{p}} \sum_{g\in\mathcal{E}} \eta'_g\|\mathbf{p}_g\|_2 + \frac{1}{2\lambda'_t}\|\mathbf{p}-\mathbf{p}'^{(t-1)}\|_2^2 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\mathbf{s}'^{(t-1)} + \mathbf{M}\mathbf{p}\|_2 \le \epsilon, \qquad (12a)\\ \mathbf{s}'^{(t)} &= \operatorname*{argmin}_{\mathbf{s}} \|\boldsymbol{\Psi}\mathbf{s}\|_0 + \frac{1}{2\mu'_t}\|\mathbf{s}-\mathbf{s}'^{(t-1)}\|_2^2 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\mathbf{s} + \mathbf{M}\mathbf{p}'^{(t)}\|_2 \le \epsilon,\ (\mathbf{e}_1\otimes\mathbf{1}_{O-1})^\top\mathbf{s} = 1, \qquad (12b)\end{aligned}$$

and (5') with

$$\begin{aligned}\hat{\mathbf{p}}^{(t)} &= \operatorname*{argmin}_{\mathbf{p}} \sum_{g\in\mathcal{E}} \eta'_g\|\mathbf{p}_g\|_2 + \frac{1}{2\hat{\lambda}_t}\|\mathbf{p}-\hat{\mathbf{p}}^{(t-1)}\|_2^2 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\hat{\mathbf{s}}^{(t-1)} + \mathbf{M}\mathbf{p}\|_2 \le \epsilon, \qquad (13a)\\ \hat{\mathbf{s}}^{(t)} &= \operatorname*{argmin}_{\mathbf{s}} \|\boldsymbol{\Psi}\mathbf{s}\|_1 + \frac{1}{2\hat{\mu}_t}\|\mathbf{s}-\hat{\mathbf{s}}^{(t-1)}\|_2^2 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\mathbf{s} + \mathbf{M}\hat{\mathbf{p}}^{(t)}\|_2 \le \epsilon,\ (\mathbf{e}_1\otimes\mathbf{1}_{O-1})^\top\mathbf{s} = 1, \qquad (13b)\end{aligned}$$

for $t \in \mathbb{N}$, where the parameters λ'_t, μ'_t, $\hat{\lambda}_t$, and $\hat{\mu}_t$ are bounded above and below by positive real numbers. By the proximal term in each update step of (12) and (13), the subproblems are strongly convex, and thus each iteration has a unique solution. Furthermore, for every $t \in \mathbb{N}$ and any given pair of constants $C_t^s, C_t^p > 0$, we may select positive values λ'_t, μ'_t, $\hat{\lambda}_t$, and $\hat{\mu}_t$ such that the solutions to (12) and (13) are equivalent to

$$\begin{aligned}\mathbf{p}'^{(t)} &= \operatorname*{argmin}_{\mathbf{p}} \sum_{g\in\mathcal{E}} \eta'_g\|\mathbf{p}_g\|_2 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\mathbf{s}'^{(t-1)} + \mathbf{M}\mathbf{p}\|_2 \le \epsilon,\ \|\mathbf{p}-\mathbf{p}'^{(t-1)}\|_2 \le C_t^p, \qquad (14a)\\ \mathbf{s}'^{(t)} &= \operatorname*{argmin}_{\mathbf{s}} \|\boldsymbol{\Psi}\mathbf{s}\|_0 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\mathbf{s} + \mathbf{M}\mathbf{p}'^{(t)}\|_2 \le \epsilon,\ (\mathbf{e}_1\otimes\mathbf{1}_{O-1})^\top\mathbf{s} = 1 \\ &\qquad\quad \|\mathbf{s}-\mathbf{s}'^{(t-1)}\|_2 \le C_t^s, \qquad (14b)\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{p}}^{(t)} &= \operatorname*{argmin}_{\mathbf{p}} \sum_{g\in\mathcal{E}} \eta'_g\|\mathbf{p}_g\|_2 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\hat{\mathbf{s}}^{(t-1)} + \mathbf{M}\mathbf{p}\|_2 \le \epsilon,\ \|\mathbf{p}-\hat{\mathbf{p}}^{(t-1)}\|_2 \le C_t^p, \qquad (15a)\\ \hat{\mathbf{s}}^{(t)} &= \operatorname*{argmin}_{\mathbf{s}} \|\boldsymbol{\Psi}\mathbf{s}\|_1 \\ &\qquad \text{s.t. } \|\boldsymbol{\Sigma}\mathbf{s} + \mathbf{M}\hat{\mathbf{p}}^{(t)}\|_2 \le \epsilon,\ (\mathbf{e}_1\otimes\mathbf{1}_{O-1})^\top\mathbf{s} = 1 \\ &\qquad\quad \|\mathbf{s}-\hat{\mathbf{s}}^{(t-1)}\|_2 \le C_t^s. \qquad (15b)\end{aligned}$$

Let us initialize the proximal alternating minimization steps for (14) and (15) with $\mathbf{p}_0 := \mathbf{p}'^{(0)} = \hat{\mathbf{p}}^{(0)}$ and $\mathbf{s}_0 := \mathbf{s}'^{(0)} = \hat{\mathbf{s}}^{(0)}$ such that $\|\boldsymbol{\Sigma}\mathbf{s}_0 + \mathbf{M}\mathbf{p}_0\|_2 < \epsilon$. Note that the objective functions of (3') and (5') are semi-algebraic functions [48] and thus have the Kurdyka-Łojasiewicz property [46]. By [46, Theorem 3.3], there exist constants $r', s' > 0$ such that when we let $\|\mathbf{p}'-\mathbf{p}_0\|_2 + \|\mathbf{s}'-\mathbf{s}_0\|_2 < r'$ and

$$\begin{aligned}\|\boldsymbol{\Psi}\mathbf{s}'\|_0 + \sum_{g\in\mathcal{E}}\eta'_g\|\mathbf{p}'_g\|_2 &\le \|\boldsymbol{\Psi}\mathbf{s}_0\|_0 + \sum_{g\in\mathcal{E}}\eta'_g\|[\mathbf{p}_0]_g\|_2 \\ &< \|\boldsymbol{\Psi}\mathbf{s}'\|_0 + \sum_{g\in\mathcal{E}}\eta'_g\|\mathbf{p}'_g\|_2 + s',\end{aligned}$$

where the first inequality is due to the optimality of $\{\mathbf{s}', \mathbf{p}'\}$, then we have that the sequence $\{\mathbf{s}'^{(t)}, \mathbf{p}'^{(t)}\}$ converges to $\{\mathbf{s}', \mathbf{p}'\}$ in finitely many steps. Similarly, there exist constants $\hat{r}, \hat{s} > 0$ such that we can guarantee that the sequence $\{\hat{\mathbf{s}}^{(t)}, \hat{\mathbf{p}}^{(t)}\}$ converges to $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$ in finitely many steps. More specifically, there exist positive integers T_1, T_2 such that

$\{\mathbf{s}', \mathbf{p}'\} = \{\mathbf{s}'^{(t)}, \mathbf{p}'^{(t)}\}$ for every $t \geq T_1$ and $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\} = \{\hat{\mathbf{s}}^{(t)}, \hat{\mathbf{p}}^{(t)}\}$ for every $t \geq T_2$.

We first show that $\mathbf{p}'^{(t)} = \hat{\mathbf{p}}^{(t)}$ for every $t \in \mathbb{N}$. Let $T := \max\{T_1, T_2\}$. Furthermore, let us consider sequences of positive real numbers C_t^s, C_t^p for $t = 1, \ldots, T$ such that

$$\sum_{t=1}^{T-1} C_t^s \leq \frac{\epsilon - \|\mathbf{\Sigma s}_0 + \mathbf{M p}_0\|_2}{\sigma_{\max}(\mathbf{\Sigma})}, \tag{16a}$$

$$\sum_{t=1}^{T} C_t^p \leq \frac{\epsilon - \|\mathbf{\Sigma s}_0 + \mathbf{M p}_0\|_2 - \sigma_{\max}(\mathbf{\Sigma}) \sum_{t=1}^{T-1} C_t^s}{\sigma_{\max}(\mathbf{M})}, \tag{16b}$$

$$C_T^s \geq (2\epsilon + \sigma_{\max}(\mathbf{M}) C_T^p)/\sigma_{\min}(\mathbf{\Sigma}). \tag{16c}$$

Note that when $\mathbf{p}'^{(0)} = \hat{\mathbf{p}}^{(0)}$ and $\mathbf{s}'^{(0)} = \hat{\mathbf{s}}^{(0)}$, we have that the optimization subproblems (12a) and (13a) are equivalent, so $\mathbf{p}_1 := \mathbf{p}'^{(1)} = \hat{\mathbf{p}}^{(1)}$. Next, assume that for some $t \leq T$, we have that $\mathbf{p}'^{(l)} = \hat{\mathbf{p}}^{(l)} =: \mathbf{p}_l$ for every $l = 1, \ldots, t-1$. Then, by (16a) and (16b) we have that

$$\begin{aligned}
\|\mathbf{\Sigma s}'^{(t-1)} + \mathbf{M}\hat{\mathbf{p}}^{(t)}\|_2 \;\leq\; & \|\mathbf{\Sigma s}_0 + \mathbf{M p}_0\|_2 \\
& + \textstyle\sum_{i=1}^{t-1} \|\mathbf{\Sigma}(\mathbf{s}'^{(i)} - \mathbf{s}'^{(i-1)})\|_2 \\
& + \textstyle\sum_{i=1}^{t} \|\mathbf{M}(\hat{\mathbf{p}}^{(i)} - \hat{\mathbf{p}}^{(i-1)})\|_2 \\
\leq\; & \|\mathbf{\Sigma s}_0 + \mathbf{M p}_0\|_2 \\
& + \sigma_{\max}(\mathbf{\Sigma}) \textstyle\sum_{i=1}^{t-1} C_i^s \\
& + \sigma_{\max}(\mathbf{M}) \textstyle\sum_{i=1}^{t} C_i^p \\
\leq\; & \|\mathbf{\Sigma s}_0 + \mathbf{M p}_0\|_2 \\
& + \sigma_{\max}(\mathbf{\Sigma}) \textstyle\sum_{i=1}^{T-1} C_i^s \\
& + \sigma_{\max}(\mathbf{M}) \textstyle\sum_{i=1}^{T} C_i^p \\
\leq\; & \epsilon,
\end{aligned}$$

and by an analogous proof, we have that

$$\|\mathbf{\Sigma}\hat{\mathbf{s}}^{(t-1)} + \mathbf{M}\mathbf{p}'^{(t)}\|_2 \leq \epsilon.$$

Then $\mathbf{p}'^{(t)}$ is a feasible solution for (15a), and $\hat{\mathbf{p}}^{(t)}$ is a feasible solution for (14a). Since the solutions are unique and the objective functions are equivalent, we have that $\mathbf{p}'^{(t)} = \hat{\mathbf{p}}^{(t)} =: \mathbf{p}_t$. Thus by induction, we have that $\mathbf{p}'^{(t)} = \hat{\mathbf{p}}^{(t)}$ for every $t \in \mathbb{N}$ and $\mathbf{p}' = \hat{\mathbf{p}} = \mathbf{p}_T$.

Next we show that the solutions $\mathbf{s}'$ and $\hat{\mathbf{s}}$ are equivalent. By (16c) we have that

$$\begin{aligned}
\|\mathbf{s}'^{(T)} - \hat{\mathbf{s}}^{(T-1)}\|_2 \leq\; & \sigma_{\min}^{-1}(\mathbf{\Sigma}) \|\mathbf{\Sigma}(\mathbf{s}'^{(T)} - \hat{\mathbf{s}}^{(T-1)})\|_2 \\
\leq\; & \sigma_{\min}^{-1}(\mathbf{\Sigma}) \|\mathbf{\Sigma s}'^{(T)} + \mathbf{M p}_T\|_2 \\
& + \sigma_{\min}^{-1}(\mathbf{\Sigma}) \|\mathbf{\Sigma}\hat{\mathbf{s}}^{(T-1)} + \mathbf{M p}_{T-1}\|_2 \\
& + \sigma_{\min}^{-1}(\mathbf{\Sigma}) \|\mathbf{M}(\mathbf{p}_T - \mathbf{p}_{T-1})\|_2 \\
\leq\; & 2\sigma_{\min}^{-1}(\mathbf{\Sigma})\epsilon + (\sigma_{\max}(\mathbf{M})/\sigma_{\min}(\mathbf{\Sigma})) C_T^p \\
\leq\; & C_T^s,
\end{aligned}$$

and similarly

$$\|\hat{\mathbf{s}}^{(T)} - \mathbf{s}'^{(T-1)}\|_2 \leq C_T^s.$$

Thus, $\mathbf{s}' = \mathbf{s}'^{(T)}$ and $\hat{\mathbf{s}} = \hat{\mathbf{s}}^{(T)}$ are both feasible solutions of (14b) and (15b) at iteration T, so we may rewrite (14b) and (15b) at iteration T as

$$\begin{aligned}
\mathbf{s}' = \operatorname*{argmin}_{\mathbf{s}} \; & \|\mathbf{\Psi s}\|_0 \\
& \text{s.t.} \; \|\mathbf{\Sigma s} + \mathbf{M p}_T\|_2 \leq \epsilon, \; (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{s} = 1, \\
& \quad \|\mathbf{s} - \mathbf{s}'^{(T-1)}\|_2 \leq C_T^s, \|\mathbf{s} - \hat{\mathbf{s}}^{(T-1)}\|_2 \leq C_T^s, \quad (17) \\
\hat{\mathbf{s}} = \operatorname*{argmin}_{\mathbf{s}} \; & \|\mathbf{\Psi s}\|_1 \\
& \text{s.t.} \; \|\mathbf{\Sigma s} + \mathbf{M p}_T\|_2 \leq \epsilon, \; (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{s} = 1, \\
& \quad \|\mathbf{s} - \mathbf{s}'^{(T-1)}\|_2 \leq C_T^s, \|\mathbf{s} - \hat{\mathbf{s}}^{(T-1)}\|_2 \leq C_T^s. \quad (18)
\end{aligned}$$

Now we provide the conditions for $\mathbf{s}' = \hat{\mathbf{s}}$. We introduce a modification to the problems (17) and (18) that are parameterized by the positive real number $r > 0$ as

$$\mathbf{s}'_r = \operatorname*{argmin}_{\mathbf{s}} \|\mathbf{\Psi s}\|_0 \,\text{s.t.}\, \|\mathbf{\Phi}_r \mathbf{s} + \mathbf{R p}_T - \mathbf{b}_r\|_2 \leq \epsilon, \tag{19}$$

$$\hat{\mathbf{s}}_r = \operatorname*{argmin}_{\mathbf{s}} \|\mathbf{\Psi s}\|_1 \,\text{s.t.}\, \|\mathbf{\Phi}_r \mathbf{s} + \mathbf{R p}_T - \mathbf{b}_r\|_2 \leq \epsilon, \tag{20}$$

where we define block conformal matrices $\mathbf{\Phi}_r$ and $\mathbf{R}$ and block conformal vector $\mathbf{b}_r$ as

$$\begin{aligned}
\mathbf{\Phi}_r &= [\mathbf{\Sigma}^\top, r(\mathbf{e}_1 \otimes \mathbf{1}_{O-1}), \epsilon(C_T^s)^{-1}(\mathbf{1}_2^\top \otimes \mathbf{I}_{KO(O-1)/2})]^\top, \\
\mathbf{R} &= [\mathbf{M}^\top, \mathbf{0}_{KO^2}, \mathbf{0}_{KO^2 \times KO(O-1)}]^\top, \\
\mathbf{b}_r &= [\mathbf{0}_{KO^2}^\top, r, \epsilon(C_T^s)^{-1}\mathbf{s}'^{(T-1)\top}, \epsilon(C_T^s)^{-1}\hat{\mathbf{s}}^{(T-1)\top}]^\top. \quad (21)
\end{aligned}$$

Note that as r increases, we recover the solutions to the unmodified problems (17) and (18), where $\mathbf{s}'_r \to \mathbf{s}'$ and $\hat{\mathbf{s}}_r \to \hat{\mathbf{s}}$ as $r \to \infty$.

By the proof of Theorem 1 in [23] and Theorem 1 of [47], if $[\mathbf{\Phi}_r]_{\cdot,\mathcal{I}}$ is full column rank and there exists a positive constant $\psi > 0$ such that

$$\|\mathbf{\Psi}_{\mathcal{J}^c,\cdot}(\psi^{-2}\mathbf{\Phi}_r^\top \mathbf{\Phi}_r + \mathbf{\Psi}_{\mathcal{J}^c,\cdot}^\top \mathbf{\Psi}_{\mathcal{J}^c,\cdot})^{-1} \mathbf{\Psi}_{\mathcal{J},\cdot}^\top\|_\infty < 1 \tag{22}$$

when $r \to \infty$, then we have that $\mathbf{s}' = \hat{\mathbf{s}}$. Under condition *1)* in the statement of Theorem 1, we have that $\mathbf{\Sigma}_{\cdot,\mathcal{I}}$ is full column rank, and since $\mathbf{\Phi}_r$ consists of rows appended to $\mathbf{\Sigma}$, then $[\mathbf{\Phi}_r]_{\cdot,\mathcal{I}}$ is also full column rank. Thus, we need only show that condition *2)* implies (22) for $r \to \infty$.

By the definition of $\mathbf{\Phi}_r$ and the Sherman-Morrison formula, we have that

$$\begin{aligned}
&(\psi^{-2}\mathbf{\Phi}_r^\top \mathbf{\Phi}_r + \mathbf{\Psi}_{\mathcal{J}^c,\cdot}^\top \mathbf{\Psi}_{\mathcal{J}^c,\cdot})^{-1} \\
&\quad = \Big(\psi^{-2}(\mathbf{\Sigma}^\top \mathbf{\Sigma} + 2\epsilon^2 (C_T^s)^{-2} \mathbf{I}_{KO(O-1)/2}) + \mathbf{\Psi}_{\mathcal{J}^c,\cdot}^\top \mathbf{\Psi}_{\mathcal{J}^c,\cdot} \\
&\qquad\qquad + r^2\psi^{-2}(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top\Big)^{-1} \\
&\quad = \mathbf{T}_1 - \frac{r^2\psi^{-2}\mathbf{T}_1(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{T}_1}{1 + r^2\psi^{-2}(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{T}_1 (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})},
\end{aligned}$$

and as $r \to \infty$, we have

$$\begin{aligned}
&\lim_{r\to\infty} (\psi^{-2}\mathbf{\Phi}_r^\top \mathbf{\Phi}_r + \mathbf{\Psi}_{\mathcal{J}^c,\cdot}^\top \mathbf{\Psi}_{\mathcal{J}^c,\cdot})^{-1} \\
&\qquad = \mathbf{T}_1 - \frac{\mathbf{T}_1(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{T}_1}{(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{T}_1 (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})} \\
&\qquad = \mathbf{T}_1 - \mathbf{T}_2.
\end{aligned}$$

Then the inequality $\|\mathbf{\Psi}_{\mathcal{J}^c,\cdot}(\mathbf{T}_1 - \mathbf{T}_2)\mathbf{\Psi}_{\mathcal{J},\cdot}^\top\|_\infty < 1$ is equivalent to the condition (22) when $r \to \infty$. Thus, we have that the conditions hold for $\mathbf{s}' = \hat{\mathbf{s}}$ by Theorem 1 of [23] and Theorem 1 of [47], as desired.

Appendix B
Proof of Theorem 2

To establish an upper bound on the estimation error of (5), we first provide the following lemma necessary to determine an upper bound on the error of (5).

Lemma 1. *Under the following four conditions,*

1) $K = o(\log O)$*;*
2) $R_1 \asymp R_2 \asymp \cdots \asymp R_K$*;*
3) $\log O = o(\min\{R/(K^7(\log R)^2), (R/K^7)^{1/3}\})$*; and*
4) $\epsilon_R \geq CO\omega\sqrt{(K\log O)/R}$ *for some constant* $C > 0$*;*

with probability at least $1 - e^{-C_1 \log O}$ *for some constant* C_1 *we have that*

$$\sum_{k=1}^{K} \left\| (\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} - \mathbf{C}_{\mathcal{O}}^{(k)})\mathbf{S}_{\mathcal{O}}^{*(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}(\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} - \mathbf{C}_{\mathcal{O}}^{(k)}) \right\|_F^2 \leq \epsilon_R^2.$$

Proof. The proof of Lemma 1 follows from the proof of Claim 2 in [23]. □

Recall that $\mathbf{s}^*$ is the vectorization of the true GSO submatrices $\{\mathbf{S}_{\mathcal{O}}^{*(k)}\}_{k=1}^K$ as in (6). We show that $\{\mathbf{s}^*, \hat{\mathbf{p}}\}$ is a feasible solution to (5'). We demonstrate an upper bound on the commutativity of sample covariance submatrices and true subnetworks as

$$\begin{aligned}
&\left| \sum_{k=1}^{K} \|\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\mathbf{S}_{\mathcal{O}}^{*(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} + \hat{\mathbf{P}}^{(k)} - (\hat{\mathbf{P}}^{(k)})^\top\|_F^2 \right|^{\frac{1}{2}} \\
&\leq \left| \sum_{k=1}^{K} \left\| (\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} - \mathbf{C}_{\mathcal{O}}^{(k)})\mathbf{S}_{\mathcal{O}}^{*(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}(\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} - \mathbf{C}_{\mathcal{O}}^{(k)}) \right\|_F^2 \right|^{\frac{1}{2}} \\
&\quad + \left| \sum_{k=1}^{K} \left\| \big(\hat{\mathbf{P}}^{(k)} - (\hat{\mathbf{P}}^{(k)})^\top\big) - \big(\mathbf{P}^{*(k)} - (\mathbf{P}^{*(k)})^\top\big) \right\|_F^2 \right|^{\frac{1}{2}} \\
&\leq \epsilon_R + \alpha, \qquad (23)
\end{aligned}$$

where we have used Lemma 1, the definition of α, and the relationship in (2). Because $\sum_{j=1}^{O}[\mathbf{S}_{\mathcal{O}}^{*(k)}]_{j1} = 1$ by definition, (23) is equivalent to

$$\|\mathbf{\Sigma}\mathbf{s}^* + \mathbf{M}\hat{\mathbf{p}}\|_2 \leq \epsilon_R + \alpha = \epsilon, \qquad (24)$$

so $\{\mathbf{s}^*, \hat{\mathbf{p}}\}$ is a feasible solution to (5').

We introduce a modification of (5') to combine the constraints into one inequality. Consider the following modified optimization problem that is parameterized by $r > 0$

$$\begin{aligned}
\{\hat{\mathbf{s}}_r, \hat{\mathbf{p}}_r\} = \underset{\{\mathbf{s},\mathbf{p}\}}{\operatorname{argmin}} \quad & \|\mathbf{\Psi}\mathbf{s}\|_1 + \sum_{k=1}^{K}\sum_{i=1}^{O} \gamma_k \|\mathbf{p}_{\mathcal{E}^{(k,i)}}\|_2 \\
& + \sum_{k<k'}\sum_{i=1}^{O} \eta_{k,k'} \|\mathbf{p}_{\mathcal{E}^{(k,k',i)}}\|_2 \\
\text{s.t.} \quad & \|\bar{\mathbf{\Phi}}_r\mathbf{s} + \bar{\mathbf{R}}\mathbf{p} - \bar{\mathbf{b}}_r\|_2 \leq \epsilon, \qquad (25)
\end{aligned}$$

where $\bar{\mathbf{\Phi}}_r = [\mathbf{\Sigma}^\top, r(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})]^\top$, $\bar{\mathbf{R}} = [\mathbf{M}^\top, \mathbf{0}_{KO^2}]^\top$, and $\bar{\mathbf{b}}_r = [\mathbf{0}_{KO(O-1)/2}^\top, r]^\top$. The parameter r determines the strictness of the second constraint in (5') such that when $r \to \infty$, we have that $\hat{\mathbf{s}}_r \to \hat{\mathbf{s}}$. Note that since $(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top\hat{\mathbf{s}} = 1$ and $(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top\mathbf{s}^* = 1$, then by (24) and the definition of $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$, we have that $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$ and $\{\mathbf{s}^*, \hat{\mathbf{p}}\}$ are feasible solutions of (25) for every $r > 0$.

We next provide an upper bound on the difference between $\hat{\mathbf{s}}$ and $\mathbf{s}^*$ following the proof of Claim 1 in [23]. First, note that as in the proof of Claim 1 of [23], we have that when $\mathbf{\Sigma}$ is full column rank, then so is $\bar{\mathbf{\Phi}}_r$, which guarantees the existence of a dual certificate $\mathbf{y} = \mathbf{I}_{\mathcal{K},\cdot}^\top \operatorname{sign}(\mathbf{\Psi}_{\mathcal{K},\cdot}\mathbf{s}^*)$, where $\mathbf{\Psi}^\top\mathbf{y} = \bar{\mathbf{\Phi}}_r^\top\bar{\mathbf{\Phi}}_r(\bar{\mathbf{\Phi}}_r^\top\bar{\mathbf{\Phi}}_r)^{-1}\mathbf{\Psi}^\top\mathbf{I}_{\mathcal{K},\cdot}^\top\operatorname{sign}(\mathbf{\Psi}_{\mathcal{K},\cdot}\mathbf{s}^*) \in \operatorname{Im}(\bar{\mathbf{\Phi}}_r^\top)$, $\mathbf{y}_{\mathcal{K}} = \operatorname{sign}(\mathbf{\Psi}_{\mathcal{K},\cdot}\mathbf{s}^*)$, $\|\mathbf{y}_{\mathcal{K}^c}\|_\infty < 1$, and $\|\mathbf{\Psi}\mathbf{s}^*\|_1 = \mathbf{y}^\top\mathbf{\Psi}\mathbf{s}^*$.

Consider the following inequality

$$\|\mathbf{\Psi}\mathbf{s}^* - \mathbf{\Psi}\hat{\mathbf{s}}\|_1 \leq \|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1 + \|\mathbf{\Psi}\mathbf{s}^* - \mathbf{u}\|_1, \qquad (26)$$

where $\mathbf{u} \in \mathbb{R}^{KO(O-1)/2}$ such that $\operatorname{supp}(\mathbf{u}) \subseteq \mathcal{K}$. We derive an upper bound for the second term on the right-hand side of (26) as

$$\begin{aligned}
\|\mathbf{\Psi}\mathbf{s}^* - \mathbf{u}\|_1 &\leq \sqrt{|\mathcal{K}|}\|\mathbf{\Psi}\mathbf{s}^* - \mathbf{u}\|_2 \\
&\leq \sqrt{|\mathcal{K}|}\|\mathbf{\Psi}\mathbf{s}^* - \mathbf{\Psi}\hat{\mathbf{s}}\|_2 + \sqrt{|\mathcal{K}|}\|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1 \\
&\leq \sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})\|\mathbf{s}^* - \hat{\mathbf{s}}\|_2 \\
&\qquad + \sqrt{|\mathcal{K}|}\|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1 \\
&\leq \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})}{\sigma_{\min}(\bar{\mathbf{\Phi}}_r)}\|\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \hat{\mathbf{s}})\|_2 \\
&\qquad + \sqrt{|\mathcal{K}|}\|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1. \qquad (27)
\end{aligned}$$

For the first term on the right-hand side of (26), we have that

$$\begin{aligned}
\xi \;&:=\; \min_{\mathbf{u}:\operatorname{supp}(\mathbf{u})\subseteq\mathcal{K}} \|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1 \\
&=\; \max_{\mathbf{v}} \min_{\mathbf{u}} \|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1 \qquad (28) \\
&\qquad + \mathbf{v}^\top\mathbf{I}_{\mathcal{K}^c,\cdot}(\mathbf{u} - \mathbf{\Psi}\hat{\mathbf{s}}) + \mathbf{v}^\top\mathbf{I}_{\mathcal{K}^c,\cdot}\mathbf{\Psi}\hat{\mathbf{s}} \\
&=\; \max_{\mathbf{w}:\operatorname{supp}(\mathbf{w})\subseteq\mathcal{K}^c} \min_{\mathbf{u}} \|\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{u}\|_1 \\
&\qquad + \mathbf{w}^\top(\mathbf{u} - \mathbf{\Psi}\hat{\mathbf{s}}) + \mathbf{w}^\top\mathbf{\Psi}\hat{\mathbf{s}},
\end{aligned}$$

where (28) results from the Lagrangian of ξ and duality theory. Given the dual certificate $\mathbf{y}$, we have that

$$\begin{aligned}
\xi \;&=\; \max_{\substack{\mathbf{w}:\operatorname{supp}(\mathbf{w})\subseteq\mathcal{K}^c, \\ \|\mathbf{w}\|_\infty\leq 1}} (\mathbf{y} + \mathbf{w})^\top\mathbf{\Psi}\hat{\mathbf{s}} - \mathbf{y}^\top\mathbf{\Psi}\hat{\mathbf{s}} \\
&\leq\; \|\mathbf{\Psi}\hat{\mathbf{s}}\|_1 - \mathbf{y}^\top\mathbf{\Psi}\hat{\mathbf{s}} + \mathbf{y}^\top\mathbf{\Psi}\mathbf{s}^* - \|\mathbf{\Psi}\mathbf{s}^*\|_1 \\
&\leq\; \mathbf{y}^\top\mathbf{\Psi}(\mathbf{s}^* - \hat{\mathbf{s}}), \qquad (29)
\end{aligned}$$

where the final inequality is due to the optimality of $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$ and the feasibility of $\{\mathbf{s}^*, \hat{\mathbf{p}}\}$ for (5'). Lastly, since $\mathbf{\Psi}^\top\mathbf{y} = \bar{\mathbf{\Phi}}_r^\top\bar{\mathbf{\Phi}}_r(\bar{\mathbf{\Phi}}_r^\top\bar{\mathbf{\Phi}}_r)^{-1}\mathbf{\Psi}^\top\mathbf{I}_{\mathcal{K},\cdot}^\top\operatorname{sign}(\mathbf{\Psi}_{\mathcal{K},\cdot}\mathbf{s}^*)$, we have that

$$\begin{aligned}
&\mathbf{y}^\top\mathbf{\Psi}(\mathbf{s}^* - \hat{\mathbf{s}}) \\
&\quad\leq \operatorname{sign}(\mathbf{\Psi}_{\mathcal{K},\cdot}\mathbf{s}^*)^\top\mathbf{I}_{\mathcal{K},\cdot}\mathbf{\Psi}(\bar{\mathbf{\Phi}}_r^\top\bar{\mathbf{\Phi}}_r)^{-1}\bar{\mathbf{\Phi}}_r^\top\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \hat{\mathbf{s}}) \\
&\quad\leq \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})}{\sigma_{\min}(\bar{\mathbf{\Phi}}_r)}\|\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \hat{\mathbf{s}})\|_2, \qquad (30)
\end{aligned}$$

where the second inequality results from the fact that every positive scalar and its ℓ_2 norm are equal. We may substitute

(27) and (30) into (26) and the fact that $\mathbf{\Psi}$ is full column rank to obtain

$$\|\mathbf{s}^* - \hat{\mathbf{s}}\|_1 \leq \tau_r \|\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \hat{\mathbf{s}})\|_2,$$

where

$$\tau_r = \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})\|\mathbf{\Psi}^\dagger\|_1}{\sigma_{\min}(\bar{\mathbf{\Phi}}_r)}(2 + \sqrt{|\mathcal{K}|}). \tag{31}$$

As $r \to \infty$, we have that

$$\begin{aligned}\|\mathbf{s}^* - \hat{\mathbf{s}}\|_1 &\leq \lim_{r\to\infty} \tau_r \|\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \hat{\mathbf{s}})\|_2 \\ &\leq 2 \lim_{r\to\infty} \tau_r(\epsilon_R + \alpha),\end{aligned}$$

where by the feasibility of $\{\hat{\mathbf{s}}, \hat{\mathbf{p}}\}$ and $\{\mathbf{s}^*, \hat{\mathbf{p}}\}$ for every $r > 0$, we have that

$$\begin{aligned}\|\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \hat{\mathbf{s}})\|_2 &\leq \|\bar{\mathbf{\Phi}}_r\mathbf{s}^* + \bar{\mathbf{R}}\hat{\mathbf{p}} - \bar{\mathbf{b}}_r\|_2 \\ &\quad + \|\bar{\mathbf{\Phi}}_r\hat{\mathbf{s}} + \bar{\mathbf{R}}\hat{\mathbf{p}} - \bar{\mathbf{b}}_r\|_2 \\ &\leq 2(\epsilon_R + \alpha).\end{aligned} \tag{32}$$

Finally, we return to the equivalent matrix formulation as

$$\sum_{k=1}^{K} \|\hat{\mathbf{S}}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}\|_1 \leq 4\tau_r(\epsilon_R + \alpha). \tag{33}$$

By the end of the proof of Theorem 2 in [23], we have that $\lim_{r\to\infty} 4\tau_r \leq \tau$, as desired.

Appendix C
Proof of Corollary 1

Consider the following optimization problem

$$\begin{aligned}\min_{\{\mathbf{S}_{\mathcal{O}}^{(k)}\}_{k=1}^K} \; & \sum_{k=1}^{K} \alpha_k \|\mathbf{S}_{\mathcal{O}}^{(k)}\|_1 + \sum_{k<k'} \beta_{k,k'} \|\mathbf{S}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{(k')}\|_1 \\ \text{s.t. } & \textstyle\sum_{k=1}^K \|\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\mathbf{S}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{(k)}\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\|_F^2 \leq \epsilon_R^2, \\ & \mathbf{S}_{\mathcal{O}}^{(k)} = (\mathbf{S}_{\mathcal{O}}^{(k)})^\top, \; \operatorname{diag}(\mathbf{S}_{\mathcal{O}}^{(k)}) = \mathbf{0}, \; \forall k = 1, \ldots, K, \\ & \textstyle\sum_j [\mathbf{S}_{\mathcal{O}}^{(1)}]_{j1} = 1,\end{aligned} \tag{34}$$

whose solution is equivalent to the naive solution $\{\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)}\}_{k=1}^K$ described in the statement of Corollary 1. Similarly to (5), we can define a vectorized version of (34) as

$$\tilde{\mathbf{s}} = \operatorname*{argmin}_{\mathbf{s}} \|\mathbf{\Psi}\mathbf{s}\|_1 \; \text{s.t.} \; \|\mathbf{\Sigma}\mathbf{s}\|_2 \leq \epsilon_R, \; (\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top \mathbf{s} = 1, \tag{35}$$

and a version parameterized by $r > 0$ as

$$\tilde{\mathbf{s}}_r = \operatorname*{argmin}_{\mathbf{s}} \|\mathbf{\Psi}\mathbf{s}\|_1 \; \text{s.t.} \; \|\bar{\mathbf{\Phi}}_r\mathbf{s} - \bar{\mathbf{b}}_r\|_2 \leq \epsilon_R, \tag{36}$$

where $\bar{\mathbf{\Phi}}_r$ and $\bar{\mathbf{b}}_r$ are defined as for (25) and $\lim_{r\to\infty} \tilde{\mathbf{s}}_r = \tilde{\mathbf{s}}$.

We provide the following upper bound via (2)

$$\begin{aligned}&\left|\sum_{k=1}^{K} \|\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\mathbf{S}_{\mathcal{O}}^{*(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}\hat{\mathbf{C}}_{\mathcal{O}}^{(k)}\|_F^2\right|^{\frac{1}{2}} \\ &\leq \left|\sum_{k=1}^{K} \left\|(\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} - \mathbf{C}_{\mathcal{O}}^{(k)})\mathbf{S}_{\mathcal{O}}^{*(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}(\hat{\mathbf{C}}_{\mathcal{O}}^{(k)} - \mathbf{C}_{\mathcal{O}}^{(k)})\right\|_F^2\right|^{\frac{1}{2}} \\ &+ \left|\sum_{k=1}^{K} \left\|\mathbf{P}^{*(k)} - (\mathbf{P}^{*(k)})^\top\right\|_F^2\right|^{\frac{1}{2}} \\ &\leq \epsilon_R + \delta,\end{aligned}$$

and similarly to Theorem 2, we apply Lemma 1 to get

$$\|\bar{\mathbf{\Phi}}_r\mathbf{s}^* - \bar{\mathbf{b}}_r\|_2 \leq \epsilon_R + \delta,$$

where $\mathbf{s}^*$ may not be a feasible solution to (36). However, by the triangle inequality and the optimality of $\tilde{\mathbf{s}}_r$, there exists $\rho \in [0, 1]$ such that

$$\|\mathbf{\Psi}\tilde{\mathbf{s}}_r\|_1 - \|\mathbf{\Psi}\mathbf{s}^*\|_1 \leq \rho\|\mathbf{\Psi}\tilde{\mathbf{s}}_r - \mathbf{\Psi}\mathbf{s}^*\|_1. \tag{37}$$

In particular, let $\rho = \max\{0, (\|\mathbf{\Psi}\tilde{\mathbf{s}}_r\|_1 - \|\mathbf{\Psi}\mathbf{s}^*\|_1)/\|\mathbf{\Psi}\tilde{\mathbf{s}}_r - \mathbf{\Psi}\mathbf{s}^*\|_1\}$, where $\rho = 0$ when $\mathbf{s}^*$ is a feasible solution to (36), but otherwise, it may be possible that $\rho \in (0, 1]$. Furthermore, since $(\mathbf{e}_1 \otimes \mathbf{1}_{O-1})^\top\tilde{\mathbf{s}} = 1$, then $\tilde{\mathbf{s}}$ is a feasible solution to (36) for every $r > 0$.

We then can introduce a similar inequality to (26) as

$$\|\mathbf{\Psi}\mathbf{s}^* - \mathbf{\Psi}\tilde{\mathbf{s}}\|_1 \leq \|\mathbf{\Psi}\tilde{\mathbf{s}} - \tilde{\mathbf{u}}\|_1 + \|\mathbf{\Psi}\mathbf{s}^* - \tilde{\mathbf{u}}\|_1, \tag{38}$$

where $\tilde{\mathbf{u}} \in \mathbb{R}^{KO(O-1)/2}$ such that $\operatorname{supp}(\tilde{\mathbf{u}}) \subseteq \mathcal{K}$. The upper bound for the second term of the right-hand side of (38) can be found analogously to (27), where we have

$$\begin{aligned}\|\mathbf{\Psi}\mathbf{s}^* - \tilde{\mathbf{u}}\|_1 \leq & \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{\Psi})}{\sigma_{\min}(\bar{\mathbf{\Phi}}_r)}\|\bar{\mathbf{\Phi}}_r(\mathbf{s}^* - \tilde{\mathbf{s}}_r)\|_2 \\ &+ \sqrt{|\mathcal{K}|}\|\mathbf{\Psi}\tilde{\mathbf{s}}_r - \tilde{\mathbf{u}}\|_1.\end{aligned} \tag{39}$$

Similarly to (29) in the proof of Theorem 2, we can upper bound the first term as

$$\begin{aligned}\tilde{\xi} &:= \min_{\tilde{\mathbf{u}}:\operatorname{supp}(\tilde{\mathbf{u}})\subseteq\mathcal{K}} \|\mathbf{\Psi}\tilde{\mathbf{s}} - \tilde{\mathbf{u}}\|_1 \\ &\leq \|\mathbf{\Psi}\tilde{\mathbf{s}}\|_1 - \mathbf{y}^\top\mathbf{\Psi}\tilde{\mathbf{s}} + \mathbf{y}^\top\mathbf{\Psi}\mathbf{s}^* - \|\mathbf{\Psi}\mathbf{s}^*\|_1 \\ &\leq \mathbf{y}^\top\mathbf{\Psi}(\mathbf{s}^* - \tilde{\mathbf{s}}) + \rho\|\mathbf{\Psi}(\mathbf{s}^* - \tilde{\mathbf{s}})\|_1,\end{aligned} \tag{40}$$

where we account for the possible infeasibility of $\mathbf{s}^*$ with (37). We may combine (40), and (39) to obtain

$$\|\tilde{\mathbf{s}} - \mathbf{s}^*\|_1 \leq (\tau_r + \tau_r')(2\epsilon_R + \delta), \tag{41}$$

where τ_r is defined in (31) and we let

$$\tau_r' := \frac{\rho KO(O-1)(1 + \sqrt{|\mathcal{K}|})\sigma_{\max}(\mathbf{\Psi})\|\mathbf{\Psi}^\dagger\|_1}{2\sigma_{\min}(\bar{\mathbf{\Phi}}_r)}.$$

As with the proof of Theorem 2, we have that for $r \to \infty$,

$$\sum_{k=1}^{K} \|\tilde{\mathbf{S}}_{\mathcal{O}}^{(k)} - \mathbf{S}_{\mathcal{O}}^{*(k)}\|_1 \leq (\tau + \tau')(\epsilon_R + \tfrac{1}{2}\delta), \tag{42}$$

as desired.

Finally, the bound (10) is equivalent to the following inequality

$$\alpha^2 \leq \left(\frac{\tau'}{\tau}\right)^2 \epsilon_R^2 + \left(\frac{\tau + \tau'}{2\tau}\right)^2 \delta^2,$$

which is a sufficient condition for the upper bound in (8) to be less than the upper bound in (9).

REFERENCES

[1] S. Rey, A. Buciulea, M. Navarro, S. Segarra, and A. G. Marques, "Joint inference of multiple graphs with hidden variables from stationary graph signals," in *IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*. IEEE, 2022, pp. 5817–5821.

[2] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York, NY: Springer, 2009.

[3] O. Sporns, *Discovering the Human Connectome*. Boston, MA: MIT Press, 2012.

[4] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[5] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[6] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, 2019.

[7] F. Xia, K. Sun, S. Yu, A. Aziz, L. Wan, S. Pan, and H. Liu, "Graph learning: A survey," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 109–127, 2021.

[8] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[9] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, 2013.

[10] P. Djuric and C. Richard, *Cooperative and Graph Signal Processing: Principles and Applications*. Academic Press, 2018.

[11] S. Rey, V. M. Tenorio, and A. G. Marques, "Robust graph filter identification and graph denoising from signal observations," *arXiv preprint arXiv:2210.08488*, 2022.

[12] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, pp. 1436–1462, 2006.

[13] V. Kalofolias, "How to learn a graph from smooth signals," in *Intl. Conf. Artif. Intel. Statist. (AISTATS)*. J. Mach. Learn. Res., 2016, pp. 920–929.

[14] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, 2016.

[15] S. S. Saboksayr and G. Mateos, "Accelerated graph learning from smooth signals," *IEEE Signal Process. Lett.*, vol. 28, pp. 2192–2196, 2021.

[16] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, 2017.

[17] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, 2017.

[18] R. Shafipour and G. Mateos, "Online topology inference from streaming stationary graph signals with partial connectivity information," *Algorithms*, vol. 13, no. 9, p. 228, 2020.

[19] T. M. Roddenberry, M. Navarro, and S. Segarra, "Network topology inference with graphon spectral penalties," in *IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*. IEEE, 2021, pp. 5390–5394.

[20] S. Rey, T. M. Roddenberry, S. Segarra, and A. G. Marques, "Enhanced graph-learning schemes driven by similar distributions of motifs," *arXiv preprint arXiv:2207.04747*, 2022.

[21] Y. Murase, J. Török, H. H. Jo, K. Kaski, and J. Kertész, "Multilayer weighted social network model," *Physical Review E*, vol. 90, no. 5, p. 052810, 2014.

[22] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *J. Roy. Statistical Soc.: Ser. B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[23] M. Navarro, Y. Wang, A. G. Marques, C. Uhler, and S. Segarra, "Joint inference of multiple graphs from matrix polynomials," *J. Mach. Learn. Res.*, vol. 23, no. 76, pp. 1–35, 2022.

[24] J. Arroyo, A. Athreya, G. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein, "Inference for multiple heterogeneous networks with a common invariant subspace," *J. Mach. Learn. Res.*, vol. 22, no. 142, pp. 1–49, 2021.

[25] M. Navarro and S. Segarra, "Joint network topology inference via a shared graphon model," *IEEE Trans. Signal Process.*, 2022.

[26] Y. Wang, S. Segarra, and C. Uhler, "High-dimensional joint estimation of multiple directed Gaussian graphical models," *Elec. J. Statist.*, vol. 14, no. 1, pp. 2439–2483, 2020.

[27] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *Annu. Allerton Conf. Commun., Control, Comput.*, vol. 40, no. 4, pp. 1935–1967, 2012.

[28] A. Chang, T. Yao, and G. I. Allen, "Graphical models and dynamic latent factors for modeling functional brain connectivity," *IEEE Data Science Wrksp. (DSW)*, pp. 57–63, 2019.

[29] A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade, "Learning linear Bayesian networks with latent variables," in *Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 249–257.

[30] J. Mei and J. M. F. Moura, "SILVar: Single index latent variable models," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2790–2803, 2018.

[31] A. Buciulea, S. Rey, C. Cabrera, and A. G. Marques, "Network reconstruction from graph-stationary signals with hidden variables," in *Conf. Signals, Syst., Computers (Asilomar)*. IEEE, 2019, pp. 56–60.

[32] A. Buciulea, S. Rey, and A. G. Marques, "Learning graphs from smooth and graph-stationary signals with hidden variables," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 273–287, 2022.

[33] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 5911–5926, 2017.

[34] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graphical Statist.*, vol. 22, no. 2, pp. 231–245, 2013.

[35] S. Rey, M. Navarro, A. Buciulea, S. Segarra, and A. G. Marques, "Joint graph learning from Gaussian observations in the presence of hidden nodes," *arXiv preprint arXiv:2212.01816*, 2022.

[36] B. Pasdeloup, V. Gripon, G. Mercier, D. Pastor, and M. G. Rabbat, "Characterization and inference of graph diffusion processes from observations of stationary signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 3, pp. 481–496, 2017.

[37] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3462–3477, 2017.

[38] B. Girault, P. Gonçalves, and E. Fleury, "Translation on graphs: An isometric shift operator," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2416–2420, 2015.

[39] Y. Zhu, M. T. Schaub, A. Jadbabaie, and S. Segarra, "Network inference from consensus dynamics with unknown parameters," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 300–315, 2020.

[40] D. Thanou, X. Dong, D. Kressner, and P. Frossard, "Learning heat diffusion graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 484–499, 2017.

[41] Y. Li and G. Mateos, "Identifying structural brain networks from functional connectivity: A network deconvolution approach," in *IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2019, pp. 1135–1139.

[42] S. Segarra, A. G. Marques, and A. Ribeiro, "Optimal graph-filter design and applications to distributed linear network operators," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4117–4131, 2017.

[43] E. Isufi, F. Gama, D. I. Shuman, and S. Segarra, "Graph filters for signal processing and machine learning on graphs," *arXiv preprint arXiv:2211.08854*, 2022.

[44] S. Segarra, G. Mateos, A. G. Marques, and A. Ribeiro, "Blind identification of graph filters," *IEEE Trans. Signal Process.*, vol. 65, no. 5, pp. 1146–1159, 2017.

[45] Y. Zhu, F. J. I. Garcia, A. G. Marques, and S. Segarra, "Estimating network processes via blind identification of multiple graph filters," *IEEE Trans. Signal Process.*, vol. 68, pp. 3049–3063, 2020.

[46] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality," *Math. Operations Res.*, vol. 35, no. 2, pp. 438–457, 2010.

[47] H. Zhang, M. Yan, and W. Yin, "One condition for solution uniqueness and robustness of both l1-synthesis and l1-analysis minimizations," *Advances in Computat. Math.*, vol. 42, no. 6, pp. 1381–1399, 2016.

[48] S. Friedland and M. Stawiska, "Some approximation problems in semi-algebraic geometry," *Banach Center Publications*, vol. 107, pp. 133–147, 2015.