Journal of Machine Learning Research 23 (2022) 1-35 Submitted 12/20; Revised 1/22; Published 2/22

Joint Inference of Multiple Graphs from Matrix Polynomials

Madeline Navarro NAV@RICE.EDU
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77005-1827, USA

Yuhao Wang YUHAOW@TSINGHUA.EDU.CN
Institute for Interdisciplinary Information Sciences
Tsinghua University and Shanghai Qi Zhi Institute
Haidian District, Beijing, China

Antonio G. Marques ANTONIO.GARCIA.MARQUES@URJC.ES
Department of Signal Theory and Communications
King Juan Carlos University
Madrid, Spain

Caroline Uhler CUHLER@MIT.EDU
Department of Electrical Engineering & Computer Science
Department of Institute for Data, Systems, and Society
Massachusetts Institute of Technology Cambridge, MA 02139-4301, USA

Santiago Segarra SEGARRA@RICE.EDU
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77005-1827, USA

Editor: Garvesh Raskutti

Abstract

Inferring graph structure from observations on the nodes is an important and popular network science task. Departing from the more common inference of a single graph, we study the problem of jointly inferring multiple graphs from the observation of signals at their nodes (graph signals), which are assumed to be stationary in the sought graphs. Graph stationarity implies that the mapping between the covariance of the signals and the sparse matrix representing the underlying graph is given by a matrix polynomial. A prominent example is that of Markov random fields, where the inverse of the covariance yields the sparse matrix of interest. From a modeling perspective, stationary graph signals can be used to model linear network processes evolving on a set of (not necessarily known) networks. Leveraging that matrix polynomials commute, a convex optimization method along with sufficient conditions that guarantee the recovery of the true graphs are provided when perfect covariance information is available. Particularly important from an empirical viewpoint, we provide high-probability bounds on the recovery error as a function of the number of signals observed and other key problem parameters. Numerical experiments demonstrate the effectiveness of the proposed method with perfect covariance information as well as its robustness in the noisy regime.

Keywords: Network topology inference, graph signal processing, spectral graph theory, multi-layer graphs, network diffusion processes.

1. Introduction

Inferring the topology of a network (graph) from a set of nodal observations is a prominent problem in statistics, network science, machine learning, and signal processing (SP) (Kolaczyk 2009; Sporns 2012), with applications including power, communications, and brain networks (Ortega et al. 2018; Marques et al. 2020; Djuric and Richard 2018), to name a few. Networks can exist as actual physical entities or can be convenient mathematical *representations* describing *parsimonious pairwise relationships* between data. Transversal to the particularities of the setup, the fundamental assumption in these network-inference approaches is the formalization of a relation between the topology of the sought network and the properties of the nodal observations. Notable approaches include correlation networks (Kolaczyk 2009, Ch. 7.3.1), partial correlations and (Gaussian) Markov random fields (Meinshausen and Bühlmann 2006; Friedman et al. 2008; Kolaczyk 2009; Lake and Tenenbaum 2010; Yuan and Lin 2007; Banerjee et al. 2008), structural equation models (Cai et al. 2013; Baingana et al. 2014), graph-SP-based approaches (Mateos et al. 2019; Dong et al. 2019; Mei and Moura 2015; Dong et al. 2016; Kalofolias 2016; Pavez and Ortega 2016; Segarra et al. 2017a; Pasdeloup et al. 2017), as well as their non-linear generalizations (Karanikolas et al. 2016; Shen et al. 2017).

While most of the existing works have looked at the problem of identifying a single network, many contemporary setups involve *multiple* related networks, each of them with a subset of available observations. Examples of this multi-graph setup arise in multi-hop communication networks deployed in *dynamic* environments where links are created or destroyed as nodes change their position, in brain analytics where observations for different patients are available and the objective is to estimate their brain functional networks, in gene-to-gene networks where the goal is to identify pairwise interactions between genes and measurements for different tissues, or in social networks where the same set of users can have different types of social interactions (Arroyo et al. 2021; Bindu et al. 2017; Murase et al. 2014; Ricchi et al. 2021).

Arguably, in many contemporary applications, dealing with multiple networks may be more the rule than the exception. Last but not least, one must also note that the joint identification of multiple graphs can be useful even if the interest is only in one of the networks, since joint formulations exploit additional sources of information and, hence, are likely to give rise to better solutions.

Given the previous motivations, our goal in this paper is to develop new *schemes for the joint inference of multiple networks* that build on recent results from graph SP (GSP) and, in particular, on the notion of graph stationarity (Marques et al. 2017; Perraudin and Vandergheynst 2017; Girault et al. 2015). In the last years, GSP has emerged as a way to generalize tools originally conceived to process signals with regular supports (time or space) to signals defined in heterogeneous domains represented by graphs (Ortega et al. 2018). The systematic approach put forth relies on the definition of a *graph shift operator* (GSO), which is a *sparse square matrix* capturing the local interactions (connections) between pairs of nodes. Within the GSP framework, the GSO constitutes the basic signal operator in the vertex domain, and its eigenvectors define the *graph* Fourier transform, which enables the analysis and processing of graph signals in a proper frequency domain. The GSO (typically assumed to have the form of an adjacency or Laplacian matrix) is also critical to define

the notion of graph stationarity (Marques et al. 2017; Perraudin and Vandergheynst 2017; Girault et al. 2015), which generalizes the classical notion of time-stationarity to signals defined on graphs and constitutes the fundamental GSP concept utilized in this paper. Given the covariance matrix associated with a random graph process, *graph stationarity* requires this covariance and the GSO representing the support of the process to have the same eigenvectors. This requirement, which is equivalent to saying that there exists a *polynomial mapping* between the *sparse shift* and the *covariance* matrix, is fairly general, encompassing classical approaches such as correlation and conditional independent networks (Mateos et al. 2019).

Leveraging those concepts, we can now describe more concretely the GSP-inspired approach put forth in this paper, which aims at inferring the topology of the multiple networks by solving an optimization problem where we look for graph shift matrices that are sparse, guarantee that the observed signals are stationary on the identified graphs, and force the different shifts to be close to each other according to a pre-specified level of similarity. Our formulation also takes into account additional structural information that may be available (such as the GSOs corresponding to a particular type of Laplacian, or being an adjacency matrix without self-loops). Together with the novel approach for the formulation of the joint topology inference of multiple networks, the paper also identifies theoretical conditions under which convex relaxations are able to find the optimal sparse structure in noiseless settings (Theorem 1) as well as a detailed theoretical analysis of the probability of robust recovery in the more practical noisy scenario (Theorem 2). [1]

1.1 Related Work

Although noticeably less than its single-network counterpart, joint inference of multiple networks—a structure oftentimes referred to as a multi-layer graph (Oselio et al. 2014; Sardellitti et al. 2019)—has attracted attention for different versions of the problem. The most widely studied one is that of inferring (tracking) the topology of time-varying networks. The standard approach is to assume that the variation is smooth across time, so that the graph-inference problem is regularized with a term that promotes changes between consecutive graphs to be small in some pre-specified norm (Baingana and Giannakis 2017; Ha et al. 2021; Kalofolias et al. 2017; Kao et al. 2017; Natali et al. 2021; Sardellitti et al. 2021; Yamada and Tanaka 2021; Yamada et al. 2019; Zhou et al. 2010).

A second cluster of works focuses on the joint inference of multiple Gaussian graphical models, where conditional dependence among nodes for a particular graphical model is represented by the precision matrix of a multivariate Gaussian distribution (Lauritzen 1996). Each graph has its own subset of signal observations, and the goal is the joint recovery of sparse precision matrices (Bilgrau et al. 2020; Cai et al. 2016; Chiquet et al. 2011; Danaher et al. 2014; Gan et al. 2019; Guo et al. 2011; Honorio and Samaras 2010; Ma and Michailidis 2016; Mohan et al. 2014; Peeters et al. 2020; Price et al. 2021; Ryali et al. 2012; Tao et al.

1. This paper significantly expands on our conference precursor (Segarra et al. 2017). The precursor simply presents the joint topology inference method for noiseless and robust settings and introduces but does not prove Theorem 1. In contrast, here we prove Theorem 1, we introduce and prove our main theoretical result (Theorem 2), we significantly enrich the discussion around our findings, and expand the numerical experiments to illustrate the new results and to include real-world data.

2016; Varoquaux et al. 2010; Wang et al. 2020; Yang et al. 2015, 2021; Zhu and Li 2018; Zhu et al. 2014).

The formulated problems typically correspond to generalizations of the graphical lasso formulation, a popular choice for estimation of Gaussian graphical models in the single graph case (Meinshausen and Bühlmann 2006; Friedman et al. 2008; Kolaczyk 2009; Lake and Tenenbaum 2010; Yuan and Lin 2007; Banerjee et al. 2008; Kumar et al. 2020). These previous works form assumptions for either similarity or common structure across the multiple graphs to improve inference of the precision matrices or similar graph structure representations. For example, Ma and Michailidis (2016) estimate multiple graphical models by first estimating all structures by group lasso given structural relationship assumptions, then applying graphical lasso given the previously estimated structure to refit the weights of the matrices. Differently, instead of directly inferring the precision matrices, Lee and Liu (2015) represent them as the sum of a common matrix and a matrix unique to each graph and apply an ℓ_1 minimization problem to estimate the decomposed matrices. As another example, Peterson et al. (2015) perform link estimation by Bayesian inference of graph structures given structural knowledge in the form of a Markov random field prior to encourage structural similarity in graphs that are more related. We refer the reader to the review of joint Gaussian graphical model inference by Tsai et al. 2021.

A third class of more involved approaches looks at the case of signal mixtures, where the assignment of each graph to observed signals is unknown; see, e.g. Lotsi and Wit (2016); Hao et al. (2017) for sparse precision-matrices approaches and Araghi et al. (2019); Hong and Dai (2021); Maretic and Frossard (2020) for GSP-based ones. In those cases, not only the graphs but also the signal-to-graph assignments must be inferred. This results in recovery problems that are more challenging to solve, with Gaussianity being often assumed to leverage expectation-maximization approaches. In most cases, the focus is on the problem formulation and algorithmic design, without characterizing the recovery performance theoretically. The present paper is more closely related to the second cluster of works but goes beyond sparse precision matrices and provides novel theoretical guarantees, as detailed next.

1.2 Contributions

This paper's contributions are fourfold:

(i) We propose an efficient optimization-based solution to the problem of joint inference of sparse graphs from the observation of stationary graph signals.

(ii) We determine sufficient conditions under which the proposed efficient method is guaranteed to recover the underlying set of true sparse graphs (Theorem 1).

(iii) We show the robustness of our method by deriving tight high-probability upper bounds on the recovery errors when imperfect covariance information is used to solve the joint inference problem (Theorem 2).

(iv) We rely on both synthetic and real-world data to compare the performance of joint and separate inference, validate the conditions for guaranteed recovery, and demonstrate the robustness of the proposed method in noisy settings.

1.3 Paper Outline

The remainder of this paper is organized as follows. In Section 2, we describe the main problem and the required assumptions, introduce some background on signal stationarity, and discuss graph similarity notions. Section 3 first introduces the non-convex problem of jointly inferring multiple graphs given covariance matrices. While true covariances are not often available, this problem lays the foundation for more realistic problem setups. The inference problem is further developed in Section 3.1, where we introduce the convex relaxation of the sparse graph learning problem and show conditions that lead to perfect recovery. In Section 4, we demonstrate the robustness of our method when only noisy or imperfect covariance matrices are available, and we provide a novel bound on the recovery error. Through experiments on synthetic and real-world data, we illustrate the performance of the proposed joint graph inference method in Section 5. Finally, we discuss conclusions and possible future research directions in Section 6.

1.4 Notation

The entries of a matrix $\mathbf{X}$ and a (column) vector $\mathbf{x}$ are denoted by X_{ij} and x_i, respectively. The notation $^\top$ and † stands for transpose and pseudo-inverse, respectively. With the size clear from the context, $\mathbf{0}$ and $\mathbf{1}$ refer to the all-zero and all-one vectors, and $\mathbf{e}_i$ refers to the i-th canonical vector, i.e., a vector whose entries are all zero except the i-th one, which is set to one. Sets are represented by calligraphic capital letters. Given an implicit set $\mathcal{B}$ and a set $\mathcal{A} \subseteq \mathcal{B}$, the set $\mathcal{A}^c$ stands for the complement set of $\mathcal{A}$, i.e., $\mathcal{A}^c = \mathcal{B} \setminus \mathcal{A}$ contains the elements in $\mathcal{B}$ that do not belong to $\mathcal{A}$. Moreover, $\mathbf{X}_\mathcal{I}$ denotes a submatrix of $\mathbf{X}$ formed by selecting the rows of $\mathbf{X}$ indexed by $\mathcal{I}$. The expression $\mathbf{X}_\mathcal{I}^\top$ denotes first selecting the rows and then transposing, whereas $[\mathbf{X}^\top]_\mathcal{I}$ is used to denote the opposite order of operations. For a vector $\mathbf{x}$, $\operatorname{diag}(\mathbf{x})$ is a diagonal matrix whose i-th diagonal entry is x_i; when applied to a matrix, $\operatorname{diag}(\mathbf{X})$ is a vector with the diagonal elements of $\mathbf{X}$. The vertical concatenation of the columns of $\mathbf{X}$ is denoted as $\operatorname{vec}(\mathbf{X})$. The operators $\circ$, $\otimes$, and $\odot$ stand for the Hadamard (element-wise), Kronecker, and Khatri-Rao (column-wise Kronecker) matrix products, while the operator $\oplus$ denotes the Kronecker matrix sum, so that $\mathbf{X} \oplus \mathbf{Y} = \mathbf{X} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Y}$, where the size of each of the identity matrices is chosen to make the dimensions of the matrices consistent. $\|\mathbf{X}\|_p$ is the matrix norm induced by the vector ℓ_p norm, not to be confused with $\|\operatorname{vec}(\mathbf{X})\|_p$. $\ker(\mathbf{X})$ and $\operatorname{Im}(\mathbf{X})$ refer to the null space and the span of the columns of $\mathbf{X}$, respectively. The notation $O(\cdot)$ and $o(\cdot)$ entail the usual asymptotic meaning and we write that $f \asymp g$ if $f = O(g)$ and $g = O(f)$.

1.5 Fundamentals of Graph Signal Processing

Let us consider a generic weighted and undirected graph $\mathcal{G}$ consisting of a node set $\mathcal{N}$ of known cardinality N, an edge set $\mathcal{E}$ of unordered pairs of elements in $\mathcal{N}$, and edge weights $A_{ij} \in \mathbb{R}$ such that $A_{ij} = A_{ji} \neq 0$ for all $(i,j) \in \mathcal{E}$. The edge weights A_{ij} are collected as entries of the symmetric adjacency matrix $\mathbf{A}$ and the node degrees in the diagonal matrix $\mathbf{D} := \operatorname{diag}(\mathbf{A1})$. These are used to form the combinatorial Laplacian matrix $\mathbf{L}_c := \mathbf{D} - \mathbf{A}$ and the normalized Laplacian $\mathbf{L} := \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. More broadly, one can define a generic GSO $\mathbf{S} \in \mathbb{R}^{N \times N}$ as any matrix whose off-diagonal sparsity pattern is equal to that

of the adjacency matrix of $\mathcal{G}$ (Sandryhaila and Moura 2013). Although the choice of $\mathbf{S}$ can be adapted to the problem at hand, most existing works set it to either $\mathbf{A}$, $\mathbf{L}_c$, or $\mathbf{L}$. If the GSO is symmetric, its normal eigendecomposition $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, with $\mathbf{V}$ unitary and $\mathbf{\Lambda}$ diagonal, exists. Suppose now that we associate a value (observation) with each node of the graph. Those N values form a graph signal that can be conveniently represented as the vector $\mathbf{x} = [x_1, ..., x_N]^\top \in \mathbb{R}^N$, with entry x_n denoting the signal value at node n. A key aspect when dealing with graph signals is the definition of meaningful operators able to relate different signals while efficiently accounting for the topology of the graph. Linear graph filters, which are defined as $\mathbf{H} = \sum_{l=0}^{\infty} h_l \mathbf{S}^l$, i.e., matrix polynomials of the GSO (Sandryhaila and Moura 2013), are the most widely-adopted alternative. Graph filters have shown to be useful not only to process graph signals (e.g., used for denoising and interpolation), but also to model linear network dynamics and network processes (Djuric and Richard 2018). To illustrate this latter point, consider a dynamic network setup where the initial state (value) of most nodes is zero and only a few seeding nodes (sources) have non-zero values. Suppose further that as time evolves, nodes communicate with their neighbors according to some dynamics captured by $h_0, h_1, ...$, then the resultant state $\mathbf{x}$ can be represented as $\mathbf{x} = \sum_{l=0}^{\infty} h_l \mathbf{S}^l \mathbf{z} = \mathbf{H}\mathbf{z}$, i.e., the output of a graph filter to a sparse input graph signal $\mathbf{z}$. The expression $\mathbf{x} = \sum_{l=0}^{\infty} h_l \mathbf{S}^l \mathbf{z}$ with $\|\mathbf{z}\|_0 \ll N$ has indeed been used to model a number of network dynamics as well as to solve different inverse problems involving observations of network processes (Segarra et al. 2017c; Djuric and Richard 2018; Segarra et al. 2017b; Zhu et al. 2020a,b).

1.6 Stationary Graph Signals

Consider now a statistical GSP setup where the values in $\mathbf{x}$ are random, and use $\bar{\mathbf{x}} = \mathbb{E}[\mathbf{x}]$ and $\mathbf{C} = \mathbb{E}\left[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^\top\right]$ to denote the mean and covariance of this random process. In this setup, the *random* graph process $\mathbf{x}$ is said to be stationary in the GSO $\mathbf{S}$ if its covariance matrix $\mathbf{C}$ is diagonalized by $\mathbf{V}$, the eigenvectors of the shift (Marques et al. 2017; Perraudin and Vandergheynst 2017; Girault et al. 2015). Equivalently, a *random* graph process is defined to be stationary in $\mathbf{S}$ if it can be represented as the output generated after filtering a white input with a linear graph filter $\mathbf{H} = \sum_{l=0}^{\infty} h_l \mathbf{S}^l$. Note that, when particularized to time-varying signals, the two aforementioned definitions boil down to the classical definition of stationary in time. The first definition requires stationary time processes to be uncorrelated in the Fourier domain, while the second one puts forth a generative model stating that a stationary time process can be represented as the output of a linear time-invariant filter to a white input (Marques et al. 2017). More importantly for the graph context, the second definition reveals that covariance matrices of graph-stationarity signals can be written as (positive-semidefinite) polynomials of the GSO. In other words, the set of processes that are stationary on a (sparse) GSO $\mathbf{S}$ is formed by the random processes whose covariances can be written as polynomials of $\mathbf{S}$ (Marques et al. 2017; Segarra et al. 2017a).

2. Problem Statement

To state our joint network topology inference problem, start by considering a scenario with K different graphs $\{\mathcal{G}^{(k)}\}_{k=1}^K$ defined over the same set $\mathcal{N}$ of nodes, but with possibly

different sets of edges and weights. This implies that K different GSOs $\{\mathbf{S}^{(k)}\}_{k=1}^{K}$ exist, each represented by an $N \times N$ matrix whose sparsity pattern and non-zero values may be different across k. Suppose also that, associated with each of the graphs, we have access to a set of graph signals collecting information attached to the nodes. Formally, we use matrix $\mathbf{X}^{(k)} := [\mathbf{x}_1^{(k)}, ..., \mathbf{x}_{n_k}^{(k)}] \in \mathbb{R}^{N \times n_k}$ to denote the matrix containing the n_k graph signals associated with graph $\mathcal{G}^{(k)}$. To simplify notation, we will assume that the signals are zero mean and denote the *sample* covariance of the k-th set as

$$\hat{\mathbf{C}}^{(k)} := \frac{1}{n_k} \mathbf{X}^{(k)} (\mathbf{X}^{(k)})^{\top}. \tag{1}$$

The setup that we investigate in this paper is one where the *graphs are unknown* and we want to use the *observed signals to infer their topology*. This is feasible under the assumption that the properties of the signals are related to those of the underlying graph. Intuitively, when there is no relation among the different graphs, each of the K topology inference problems can be solved separately. However, if the graphs are related, joint inference can be beneficial. In this context, our problem is stated as follows.

Problem 1 *Given the observations $\{\mathbf{X}^{(k)}\}_{k=1}^{K}$ find the graph structure encoded in $\{\mathbf{S}^{(k)}\}_{k=1}^{K}$ under the assumptions that: (AS1) the signals in $\mathbf{X}^{(k)}$ are realizations of a process that is stationary in $\mathbf{S}^{(k)}$ and (AS2) graphs k and k' are "close" according to a particular distance $d(\mathbf{S}^{(k)}, \mathbf{S}^{(k')})$.*

Although relatively formal, the statement of the problem above can give rise to different formulations. This issue will be resolved in Section 3, where an optimization problem associated with Problem 1 is presented. Before that, several remarks on assumptions (AS1) and (AS2) are provided.

(AS1) Stationarity: To better understand the implications of (AS1), let us recall that stationarity requires the covariance of the graph process to be a polynomial of $\mathbf{S}$. In other words, (AS1) is tantamount to assuming that the mapping between the GSO $\mathbf{S}$, which represents pairwise relationships between the nodes, and the matrix $\mathbf{C} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\top}\right]$, which represents pairwise correlations between the nodes, is analytic (smooth), so that it can be accurately represented by a matrix polynomial. At an intuitive level, this model assumes that $\mathbf{S}$ encodes latent *one-hop* interactions between nodes and that each successive application of the shift (i.e., higher-order powers of $\mathbf{S}$) spreads the original information across an iteratively increasing neighborhood, which ends up giving rise to *indirect* correlations among all nodes in the graph (Djuric and Richard 2018). Put it differently, although the correlation is given by the *dense* matrix $\mathbf{C}$, the actual dependencies can be (more easily) represented by the more *parsimonious* matrix $\mathbf{S}$. Relevant relations between the shift and the covariance matrices that fall within this model include

- $\mathbf{C} = \mathbf{S}$, as in correlation networks;
- $\mathbf{C} = \mathbf{S}^{-1}$, as in conditionally independent Markov random fields; or
- $\mathbf{C} = (\mathbf{I} - \mathbf{S})^{-1}(\mathbf{I} - \mathbf{S})^{-\top} = (\mathbf{I} - \mathbf{S})^{-2}$, as in symmetric structural equation models with white exogenous inputs.

To elaborate on the third example, structural equation models postulate that the observed signal $\mathbf{x}$ can be written as $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{w}$, where $\mathbf{w}$ is the so-called exogenous input, and $\mathbf{A}$ is an adjacency matrix without self-loops (Shen et al. 2017). Rewriting the previous expression as $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{w}$ and using the fact that $\mathbf{w}$ is white, it follows that $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] = (\mathbf{I}-\mathbf{A})^{-1}\mathbb{E}\left[\mathbf{w}\mathbf{w}^\top\right](\mathbf{I}-\mathbf{A})^{-\top} = (\mathbf{I}-\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})^{-\top} = (\mathbf{I}-\mathbf{A})^{-2}$, where for the last step we have used that the graph is undirected. Note also that the second example, which can be equivalently written as $\mathbf{S} = \mathbf{C}^{-1}$, will allow us to establish meaningful links between our approach and graphical lasso. Although graph stationarity does not require Gaussianity, many of the works in the area assume that the graph signals at hand are not only stationary but also Gaussian distributed (Djuric and Richard 2018). That is indeed the case for, e.g., linear network diffusion processes whose initial condition is Gaussian. While the algorithms presented in this paper can be applied regardless of the distribution of the data, the theoretical result in Theorem 2 is the only point where Gaussianity is assumed.

(AS2) Similarity Among Graphs. Regarding (AS2), the two critical issues are the form of the distance function $d(\cdot,\cdot)$ and determining the proximity degree among the different graphs. To handle the second issue, let us define the weighted and directed graph $\mathcal{G}_\mathcal{Q}$ whose node set $\mathcal{Q}$ collects the K GSOs and with $W_{k,k'}$, the weight of edge (k,k'), representing the similarity between $\mathbf{S}^{(k)}$ and $\mathbf{S}^{(k')}$. Note that this graph $\mathcal{G}_\mathcal{Q}$ is *not to be estimated* but is *user-defined* to represent the similarity relationship among the GSOs that are to be estimated $\{\mathbf{S}^{(k)}\}_{k=1}^K$. The particular form of $\mathcal{G}_\mathcal{Q}$ will depend on the application at hand. In dynamic environments where the index k corresponds to time, a reasonable choice is to set $\mathcal{G}_\mathcal{Q}$ to a *directed path* connecting the GSOs corresponding to consecutive time instants (windows). Differently, if k indexes patients with a particular disease, then it is reasonable to set $\mathcal{G}_\mathcal{Q}$ as a *complete graph* with the strength of the connection $W_{k,k'}$ depending on the similarity between the corresponding patients. The weights in $\mathcal{G}_\mathcal{Q}$ can be known beforehand or learned from the data after postulating a particular model (see, e.g., Oselio et al. (2014) for a hierarchical approach). Regarding the form of $d(\mathbf{S}^{(k)}, \mathbf{S}^{(k')})$, reasonable choices include $\|\text{vec}(\mathbf{S}^{(k)} - \mathbf{S}^{(k')})\|_0$ and $\|\text{vec}(\mathbf{S}^{(k)} - \mathbf{S}^{(k')})\|_1$, which will promote the pair of shifts to have the same sparsity pattern and weights; and $\|\text{vec}(\mathbf{S}^{(k)} - \mathbf{S}^{(k')})\|_2^2$, which will promote similar weights. Several of these distances have been explored in the context of joint identification of multiple sparse precision matrices $\mathbf{C} = \mathbf{S}^{-1}$ giving rise to modified graphical lasso formulations using regularized lasso (Danaher et al. 2014), regularized elastic net (Ryali et al. 2012), and regularized $\ell_{1,\infty}$ group lasso (Honorio and Samaras 2010).

3. Convex Solution and Recovery Guarantees

Our goal is to provide an optimization-based solution to Problem 1. Because multiple solutions satisfy the stationarity assumption (AS1), recovery includes the selection of a particular solution satisfying desirable structural characteristics. We encourage parsimonious graph structure to obtain interpretable representations and minimize computation of potential downstream tasks. Specifically, our approach is to find the sparsest graphs $\{\mathbf{S}^{(k)^*}\}_{k=1}^K$

that satisfy assumptions (AS1) and (AS2) by solving

$$\min_{\{\mathbf{S}^{(k)}\}_{k=1}^{K}} \sum_{k} \alpha_k \|\mathrm{vec}(\mathbf{S}^{(k)})\|_0 + \sum_{k<k'} \beta_{k,k'}\, d(\mathbf{S}^{(k)}, \mathbf{S}^{(k')}) \\ \text{s. t.} \quad \mathbf{C}^{(k)}\mathbf{S}^{(k)} = \mathbf{S}^{(k)}\mathbf{C}^{(k)}, \;\; \mathbf{S}^{(k)} \in \mathcal{S}^{(k)}, \;\; \text{for all } k. \tag{2}$$

In (2), the set $\mathcal{S}^{(k)}$ specifies additional properties that $\mathbf{S}^{(k)}$ must satisfy, with examples including symmetry, zero diagonal elements (if the GSO is an adjacency matrix with no self-loops), or non-positive off-diagonal elements and $\mathbf{0} = \mathbf{S}^{(k)}\mathbf{1}$ for the case of a combinatorial Laplacian. Regarding the structure of the objective, the first term promotes sparsity on the GSOs, the second one promotes the proximity postulated in (AS2), and $\{\alpha_k\}$ and $\{\beta_{k,k'}\}$ are parameters that allow a trade-off between the two terms in the objective. Finally, the constraints $\mathbf{C}^{(k)}\mathbf{S}^{(k)} = \mathbf{S}^{(k)}\mathbf{C}^{(k)}$ account for (AS1). Specifically, note that stationarity implies that the eigenvectors of the covariance and those of the GSO are the same; hence, the covariance and the shift must commute, as enforced in the constraint. As will be apparent in Section 4, to take into account that in practice we have access to the *sample* covariance $\hat{\mathbf{C}}^{(k)}$, it is reasonable to relax the equalities in $\mathbf{C}^{(k)}\mathbf{S}^{(k)} = \mathbf{S}^{(k)}\mathbf{C}^{(k)}$, with the level of tolerated violation depending on the number of samples available to form the estimates $\hat{\mathbf{C}}^{(k)}$.

3.1 Relaxation for the Sparse Formulation

Consider the following convex optimization problem

$$\min_{\{\mathbf{S}^{(k)}\}_{k=1}^{K}} \sum_{k} \alpha_k \|\mathrm{vec}(\mathbf{S}^{(k)})\|_1 + \sum_{k<k'} \beta_{k,k'} \|\mathrm{vec}(\mathbf{S}^{(k)} - \mathbf{S}^{(k')})\|_1 \\ \text{s. t.} \quad \mathbf{C}^{(k)}\mathbf{S}^{(k)} = \mathbf{S}^{(k)}\mathbf{C}^{(k)}, \;\; \mathbf{S}^{(k)} = {\mathbf{S}^{(k)}}^{\top}, \;\; \text{for all } k \\ S_{ii}^{(k)} = 0, \text{ for all } \{k, i\}, \;\; \textstyle\sum_{j=1}^{N} S_{j1}^{(1)} = 1. \tag{3}$$

Notice that from (2) to (3) we not only relax the objective function by replacing the ℓ_0-norm with the ℓ_1-norm, but we also specify the distance $d(\cdot,\cdot)$ in the objective as a graph pairwise ℓ_1-norm difference and the feasibility set $\mathcal{S}^{(k)}$ by requiring valid undirected adjacency matrices without loops. Finally, the last constraint in (3) fixes the scale of the recovered graphs and precludes the all-zero solution from belonging to the feasibility set. If we denote by $\{\hat{\mathbf{S}}^{(k)}\}_{k=1}^{K}$ the solution to (3), we now present conditions under which $\{\hat{\mathbf{S}}^{(k)}\}_{k=1}^{K}$ is guaranteed to coincide with the corresponding solution $\{{\mathbf{S}^{(k)}}^{*}\}_{k=1}^{K}$ to (2).

In order to formally define these conditions, a series of definitions must be put in place. First, define matrices $\mathbf{B}^{(i,j)} \in \mathbb{R}^{N\times N}$ for $i < j$ such that $B_{ij}^{(i,j)} = 1$, $B_{ji}^{(i,j)} = -1$, and all other entries are zero. Based on this, we denote by $\mathbf{B} \in \mathbb{R}^{\binom{N}{2}\times N^2}$ a matrix whose rows are the vectorized forms of $\mathbf{B}^{(i,j)}$ for all $i, j \in \{1, 2, \ldots, N\}$ where $i < j$. In this way, $\mathbf{B}\mathbf{s}^{(k)} = \mathbf{0}$ when $\mathbf{s}^{(k)}$ is the vectorized form of a symmetric matrix. Similarly, define vectors $\mathbf{z}^{(i,j)} \in \mathbb{R}^K$ for $i < j \leq K$ such that $z_i^{(i,j)} = 1$, $z_j^{(i,j)} = -1$, and all other entries are zero. We build the matrix $\mathbf{Z} \in \mathbb{R}^{\binom{K}{2}\times K}$ whose rows are the vectors ${\mathbf{z}^{(i,j)}}^{\top}$. We consolidate the information of all the covariances $\mathbf{C}^{(k)}$ in the block diagonal matrix $\mathbf{\Sigma}$ defined as $\mathbf{\Sigma} := \mathrm{blockdiag}(-\mathbf{C}^{(1)} \oplus \mathbf{C}^{(1)}, \ldots, -\mathbf{C}^{(K)} \oplus \mathbf{C}^{(K)})$ where, we recall, $\oplus$ denotes the Kronecker sum. With $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

collecting the values of $\{\alpha_k\}$ and $\{\beta_{k,k'}\}$ respectively, and $\mathcal{D}' = \{1, N+2, \ldots, N^2\}$ denoting the indices corresponding to the diagonal of an $N \times N$ matrix when vectorized, we define the following two matrices

$$\mathbf{\Psi} := \begin{bmatrix} \text{diag}(\boldsymbol{\alpha}) \\ \text{diag}(\boldsymbol{\beta})\mathbf{Z} \end{bmatrix} \otimes \mathbf{I}_{N^2}, \quad \mathbf{\Phi} := \begin{bmatrix} \mathbf{I}_K \otimes \mathbf{B} \\ \mathbf{I}_K \otimes [\mathbf{I}_{N^2}]_{\mathcal{D}'} \\ \mathbf{\Sigma} \\ (\mathbf{e}_1 \otimes \mathbf{1}_N)^\top \end{bmatrix}. \tag{4}$$

Denote by $\mathcal{J}$ the index set of the support of $\mathbf{s}^*$, where $\mathbf{s}^* \in \mathbb{R}^{KN^2}$ collects the vectorized versions of $\{\mathbf{S}^{(k)^*}\}_{k=1}^K$, and by $\mathcal{I}$ the index set of the support of $\mathbf{\Psi}\mathbf{s}^*$. With this notation in place, the following result holds.

Theorem 1 *Assuming problem* (3) *is feasible,* $\{\hat{\mathbf{S}}^{(k)}\}_{k=1}^K = \{\mathbf{S}^{(k)^*}\}_{k=1}^K$ *if the two following conditions are satisfied:*
1) $[\mathbf{\Phi}^\top]_{\mathcal{J}}$ *is full row rank; and*
2) There exists a constant $\delta > 0$ *such that*

$$\gamma := \|\mathbf{\Psi}_{\mathcal{I}^c}(\delta^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{\Psi}_{\mathcal{I}^c}^\top\mathbf{\Psi}_{\mathcal{I}^c})^{-1}\mathbf{\Psi}_{\mathcal{I}}^\top\|_\infty < 1. \tag{5}$$

Proof Denoting by $\mathbf{s}^{(k)} = \text{vec}(\mathbf{S}^{(k)})$ for all k, problem (3) can be reformulated as

$$\begin{aligned} \min_{\{\mathbf{s}^{(k)}\}_{k=1}^K} \quad & \sum_k \alpha_k \|\mathbf{s}^{(k)}\|_1 \; + \sum_{k<k'} \beta_{k,k'} \|\mathbf{s}^{(k)} - \mathbf{s}^{(k')}\|_1 \\ \text{s. t.} \quad & (\mathbf{I}_N \otimes \mathbf{C}^{(k)} - \mathbf{C}^{(k)} \otimes \mathbf{I}_N)\mathbf{s}^{(k)} = \mathbf{0}, \\ & \mathbf{B}\mathbf{s}^{(k)} = \mathbf{0}, \quad [\mathbf{I}_{N^2}]_{\mathcal{D}'}\, \mathbf{s}^{(k)} = \mathbf{0}, \qquad \text{for all } k \\ & (\mathbf{e}_1 \otimes \mathbf{1}_N)^\top \mathbf{s}^{(1)} = 1, \end{aligned} \tag{6}$$

where, we recall, $\mathbf{s}^{(k)}$ belonging to the null space of $\mathbf{B}$ ensures that $\mathbf{S}^{(k)}$ is symmetric, and the last equality imposes that the first column of $\mathbf{S}^{(1)}$ sums up to 1 [cf. last constraint in (3)]. Denoting by $\mathbf{s} = [\mathbf{s}^{(1)^\top}, \ldots, \mathbf{s}^{(K)^\top}]^\top$ and leveraging the definitions in (4), problem (6) can be compactly stated as

$$\min_{\mathbf{s}} \|\mathbf{\Psi}\mathbf{s}\|_1 \quad \text{s. t. } \mathbf{\Phi}\mathbf{s} = \mathbf{b}, \tag{7}$$

where $\mathbf{b}$ is a binary vector of length $K(\binom{N}{2} + N^2 + N) + 1$ with all its entries equal to 0 except for the last one that is a 1. Problem (7) is an instance of ℓ_1-analysis (Zhang et al., 2016). It can be shown through Theorem 1 by Zhang et al. (2016) that the solution to (7) coincides with the sparsest solution $\mathbf{s}^*$ if:

a) $\ker(\mathbf{\Psi}_{\mathcal{I}^c}) \cap \ker(\mathbf{\Phi}) = \{\mathbf{0}\}$; and

b) There exists a vector $\mathbf{y} \in \mathbb{R}^{N^2(K+\binom{K}{2})}$ such that $\mathbf{\Psi}^\top\mathbf{y} \in \text{Im}(\mathbf{\Phi}^\top)$, $\mathbf{y}_{\mathcal{I}} = \text{sign}(\mathbf{\Psi}_{\mathcal{I}}\mathbf{s}^*)$, and $\|\mathbf{y}_{\mathcal{I}^c}\|_\infty < 1$.

The remainder of the proof is devoted to showing that if conditions *1)* and *2)* in the statement of the theorem hold, then a) and b) are satisfied.

We begin by showing that *1)* implies a). In order to do this, we first provide some insight on the specific form of $\mathbf{\Psi}_{\mathcal{I}^c}$. Notice that the first $K\,N^2$ rows of $\mathbf{\Psi}$ correspond to the computation of the ℓ_1-norm cost of each entry of the K graph shifts whereas the last $\binom{K}{2} N^2$ rows of $\mathbf{\Psi}$ correspond to the cost of a discrepancy between corresponding entries of two different graph shifts. Hence, the rows selected in $\mathbf{\Psi}_{\mathcal{I}^c}$, i.e., the ones *not* in the support of $\mathbf{\Psi}\,\mathbf{s}^*$, belong to two classes: i) among the first $K\,N^2$ rows, $\mathcal{I}^c$ selects the rows corresponding to elements in $\mathbf{s}^*$ which are 0; and ii) among the last $\binom{K}{2} N^2$, $\mathcal{I}^c$ selects the rows corresponding to pairs of elements that are repeated in two different graph shifts. Thus, for a generic vector $\mathbf{w} \in \mathbb{R}^{K\,N^2}$ to belong to $\ker(\mathbf{\Psi}_{\mathcal{I}^c})$ two conditions must be satisfied (associated with the two aforementioned classes): i) if $s^*_i = 0$ then $w_i = 0$; and ii) if $s^*_{(k-1)N^2+i} = s^*_{(k'-1)N^2+i}$ for some k, k', i then $w_{(k-1)N^2+i} = w_{(k'-1)N^2+i}$. For a) to be satisfied, we need to guarantee that any such $\mathbf{w}$ cannot belong to the null space of $\mathbf{\Phi}$. A sufficient condition for this is to require that columns i of $\mathbf{\Phi}$ associated with values $\mathbf{s}^*_i \neq 0$ are linearly independent, which is exactly condition *1)* in the theorem's statement.

The next step is to show that condition *2)* implies b). For this, consider the following ℓ_2 norm minimization problem

$$\min_{\{\mathbf{y},\mathbf{r}\}} \delta^2 \|\mathbf{r}\|_2^2 + \|\mathbf{y}\|_2^2 \;\text{ s. t. }\; \mathbf{\Psi}^\top \mathbf{y} = \mathbf{\Phi}^\top \mathbf{r},\;\; \mathbf{y}_{\mathcal{I}} = \text{sign}(\mathbf{\Psi}_{\mathcal{I}} \mathbf{s}^*), \tag{8}$$

where δ is a positive tuning constant. Including the term $\delta^2\|\mathbf{r}\|_2^2$ in the objective guarantees the existence of a closed-form expression for the minimizing argument, while preventing numerical instability when solving the optimization. We now show that the solution $\mathbf{y}^*$ of (8) satisfies the requirements imposed in condition b). The two constraints in (8) enforce the fulfillment of the first two requirements in b), hence, we are left to show that $\|\mathbf{y}^*_{\mathcal{I}^c}\|_\infty < 1$. Since the values of $\mathbf{y}_{\mathcal{I}}$ are fixed, the constraint $\mathbf{\Psi}^\top \mathbf{y} = \mathbf{\Phi}^\top \mathbf{r}$ can be rewritten as $\mathbf{\Psi}_{\mathcal{I}}^\top \text{sign}(\mathbf{\Psi}_{\mathcal{I}} \mathbf{s}^*) = -\mathbf{\Psi}_{\mathcal{I}^c}^\top \mathbf{y}_{\mathcal{I}^c} + \mathbf{\Phi}^\top \delta^{-1} \delta \mathbf{r}$. Then, by defining the vector $\mathbf{t} := [\delta \mathbf{r}^\top, -\mathbf{y}_{\mathcal{I}^c}^\top]^\top$ and the matrix $\mathbf{Q} := [\delta^{-1}\mathbf{\Phi}^\top, \mathbf{\Psi}_{\mathcal{I}^c}^\top]$, (8) can be rewritten as

$$\min_{\mathbf{t}} \;\; \|\mathbf{t}\|_2^2 \;\text{ s. t. }\; \mathbf{\Psi}_{\mathcal{I}}^\top \text{sign}(\mathbf{\Psi}_{\mathcal{I}} \mathbf{s}^*) = \mathbf{Q}\mathbf{t}. \tag{9}$$

The minimum-norm solution to (9) is given by $\mathbf{t}^* = (\mathbf{Q})^\dagger \mathbf{\Psi}_{\mathcal{I}}^\top \text{sign}(\mathbf{\Psi}_{\mathcal{I}} \mathbf{s}^*)$ from where it follows that

$$\mathbf{y}^*_{\mathcal{I}^c} = -\mathbf{\Psi}_{\mathcal{I}^c} (\delta^{-2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{\Psi}_{\mathcal{I}^c}^\top \mathbf{\Psi}_{\mathcal{I}^c})^{-1} \mathbf{\Psi}_{\mathcal{I}}^\top \,\text{sign}(\mathbf{\Psi}_{\mathcal{I}} \mathbf{s}^*). \tag{10}$$

Condition a) guarantees the existence of the inverse in (10). Since $\|\text{sign}(\mathbf{\Psi}_{\mathcal{I}} \mathbf{s}^*)\|_\infty = 1$, we may bound the ℓ_∞ norm of $\mathbf{y}^*_{\mathcal{I}^c}$ as $\|\mathbf{y}^*_{\mathcal{I}^c}\|_\infty \leq \|\mathbf{\Psi}_{\mathcal{I}^c} (\delta^{-2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{\Psi}_{\mathcal{I}^c}^\top \mathbf{\Psi}_{\mathcal{I}^c})^{-1} \mathbf{\Psi}_{\mathcal{I}}^\top\|_\infty = \gamma$. Hence, condition *2)* in the theorem guarantees $\|\mathbf{y}^*_{\mathcal{I}^c}\|_\infty < 1$ as wanted. ■

Theorem 1 provides *sufficient* conditions under which the relaxation in (3) is guaranteed to recover the true sparse GSOs $\{\mathbf{S}^{(k)^*}\}_{k=1}^K$. Numerical experiments in Section 5 reveal that the bound imposed on γ in (5) is tight by providing examples where $\gamma = 1$ and for which

recovery fails. In the statement of the theorem, condition *1)* ensures that the solution to (3) is unique, a necessary requirement to guarantee sparse recovery. Condition *2)* is derived from the construction of a dual certificate designed to ensure that the unique solution to (3) also has minimum ℓ_0 (pseudo-)norm (Zhang et al. 2016). Details within the proof of the theorem reveal why condition *2)* is sufficient but not necessary. In a nutshell, the condition guarantees that a *specific* judicious candidate for the dual certificate (obtained by minimizing a relevant ℓ_2 norm) satisfies a bound on its ℓ_∞ norm. However, when this specific candidate fails, one cannot rule out the existence of better dual certificates that can ensure sparse recovery. To gain further intuition on (5), notice that condition *2)* is always satisfied whenever $\mathbf{\Phi}^\top\mathbf{\Phi}$ is invertible. Indeed, for small values of δ we have that $\gamma \approx \delta^2\|\mathbf{\Psi}_{\mathcal{I}^c}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Psi}_{\mathcal{I}}^\top\|_\infty$, which can be made smaller than 1 by selecting arbitrarily small values of δ. This should not be surprising since $\mathbf{\Phi}^\top\mathbf{\Phi}$ being invertible implies that $\mathbf{\Phi}$ has full column rank which, in turn, implies that the feasibility set of our problem is a singleton [cf. (7)]. Thus, in this extreme case, the ℓ_1 relaxation (and any other objective) is guaranteed to recover the true GSOs. Notice that the guarantees for exact recovery provided by Theorem 1 strongly rely on the fact that all constraints in (3) are equality constraints. This, in turn, is enabled by the assumption that we have perfect knowledge of the covariances $\mathbf{C}^{(k)}$. Thus, the more practical scenario where the covariances are estimated requires a robust reformulation of the recovery problem, as we discuss next.

4. Robust Recovery and Sample Complexity

Following the formal description of Problem 1, we do not have access to the covariance matrices $\mathbf{C}^{(k)}$ but rather to signals $\{\mathbf{X}^{(k)}\}_{k=1}^K$. Hence, we reformulate (3) to account for the fact that we can only have access to sample estimates $\hat{\mathbf{C}}^{(k)}$ of the covariances [cf. (1)]. More specifically, the commutativity constraint in (3), $\mathbf{C}^{(k)}\mathbf{S}^{(k)} = \mathbf{S}^{(k)}\mathbf{C}^{(k)}$, is relaxed and instead we bound the difference between $\hat{\mathbf{C}}^{(k)}\mathbf{S}^{(k)}$ and $\mathbf{S}^{(k)}\hat{\mathbf{C}}^{(k)}$, giving rise to the following optimization problem

$$\begin{aligned}
\min_{\{\mathbf{S}^{(k)}\}_{k=1}^K} \; & \sum_k \alpha_k \|\mathrm{vec}(\mathbf{S}^{(k)})\|_1 + \sum_{k<k'} \beta_{k,k'} \|\mathrm{vec}(\mathbf{S}^{(k)} - \mathbf{S}^{(k')})\|_1 \\
\text{s. t.} \quad & \sum_{k=1}^K \|\mathbf{S}^{(k)}\hat{\mathbf{C}}^{(k)} - \hat{\mathbf{C}}^{(k)}\mathbf{S}^{(k)}\|_{\mathrm{F}}^2 \leq \epsilon_n^2 \\
& \mathbf{S}^{(k)} = {\mathbf{S}^{(k)}}^\top, \;\; \text{for all } k \\
& S_{ii}^{(k)} = 0, \text{ for all } \{k,i\}, \;\; \textstyle\sum_{j=1}^N S_{j1}^{(1)} = 1.
\end{aligned} \tag{11}$$

Our goal is to bound the distortion between the real GSOs $\{\mathbf{S}^{(k)*}\}_{k=1}^K$ and the estimated ones $\{\hat{\mathbf{S}}^{(k)}\}_{k=1}^K$ obtained by solving (11), where ϵ_n is selected large enough to ensure feasibility. To formally state this bound, a series of definitions must be put in place.

Recalling that $\mathbf{s}^* \in \mathbb{R}^{KN^2}$ collects the vectorized versions of the true GSOs, $\{\mathbf{S}^{(k)*}\}_{k=1}^K$, we denote by $\mathcal{D}$, $\mathcal{L}$, and $\mathcal{U}$ the indices in $\mathbf{s}^*$ corresponding to the diagonal, lower triangular, and upper triangular elements of $\mathbf{S}^{(k)*}$ for $k = 1, \ldots, K$. Analogous to the definition of $\mathbf{\Sigma}$, we define the block diagonal $\hat{\mathbf{\Sigma}}$ that combines the sample covariance matrices $\hat{\mathbf{C}}^{(k)}$ as $\hat{\mathbf{\Sigma}} := \mathrm{blockdiag}(-\hat{\mathbf{C}}^{(1)} \oplus \hat{\mathbf{C}}^{(1)}, \ldots, -\hat{\mathbf{C}}^{(K)} \oplus \hat{\mathbf{C}}^{(K)})$. Define matrices $\mathbf{M} := [\hat{\mathbf{\Sigma}}]_{\mathcal{L}}^\top + [\hat{\mathbf{\Sigma}}]_{\mathcal{U}}^\top \in$

$\mathbb{R}^{KN^2\times K\binom{N}{2}}$, $\mathbf{R} := [\mathbf{\Psi}^\top]^\top_{\mathcal{L}} \in \mathbb{R}^{\left(K+\binom{K}{2}\right)N^2\times K\binom{N}{2}}$, and let $\sum_k n_k = n$. We let $\mathcal{K}$ represent the support of $\mathbf{R}\mathbf{s}^*_{\mathcal{L}}$. Finally, we define the constant $\omega := \max_{k=1,\dots,K}\omega_k$ where $\omega_k := \max\{\max_i[\mathbf{C}^{(k)}]_{ii}, \max_i[\mathbf{S}^{(k)^*}\mathbf{C}^{(k)}\mathbf{S}^{(k)^*}]_{ii}\}$. With this notation in place, we state our main result on the performance of the proposed robust recovery scheme.

Theorem 2 *If the following five conditions are satisfied:*
1) $\mathbf{M}$ *is full column rank.*
2) $K = o(\log N)$.
3) $n_1 \asymp n_2 \asymp \ldots \asymp n_K$.
4) $\log N = o(\min\{n/(K^7(\log n)^2), (n/K^7)^{1/3}\})$.
5) $\epsilon_n \geq CN\omega\sqrt{(K\log N)/n}$, *for some constant* $C > 0$.
For each graph k, *we observe graph signals as realizations of a Gaussian white noise process that is stationary in* $\mathbf{S}^{(k)}$. *Then, with probability at least* $1 - e^{-C'\log N}$ *for some constant* $C' > 0$ *we have that*

$$\begin{aligned}&\sum_{k=1}^{K}\|\mathrm{vec}(\hat{\mathbf{S}}^{(k)} - \mathbf{S}^{(k)^*})\|_1 \leq \gamma\epsilon_n,\\ &\textit{where } \gamma = \frac{4\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})\|\mathbf{R}^\dagger\|_1}{\sigma_{\min}(\mathbf{M})}\left(2+\sqrt{|\mathcal{K}|}\right).\end{aligned} \tag{12}$$

Proof We first state the following lemma, which characterizes the eigenvalues of a matrix after performing a rank-one update and that will be instrumental in showing our main result.

Lemma 1 *Golub (1973) Let* $\mathbf{C} = \mathbf{D} + \mathbf{u}\mathbf{u}^\top$ *where* $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ *is a diagonal matrix of size* $m \times m$ *such that* $d_i \leq d_{i+1}$. *We denote the eigenvalues of* $\mathbf{C}$ *by* λ_i *such that* $\lambda_i \leq \lambda_{i+1}$. *Then, for* $i = 1, \ldots, m-1$ *it holds that* $d_i \leq \lambda_i \leq d_{i+1}$.

Recalling that $\mathbf{s} = [\mathrm{vec}(\mathbf{S}^{(1)})^\top, \ldots, \mathrm{vec}(\mathbf{S}^{(K)})^\top]^\top$, we may reformulate (11) as

$$\begin{aligned}\min_{\mathbf{s}} \quad & \|(\mathrm{diag}(\boldsymbol{\alpha})\otimes\mathbf{I}_{N^2})\,\mathbf{s}\|_1 + \|(\mathrm{diag}(\boldsymbol{\beta})\mathbf{Z}\otimes\mathbf{I}_{N^2})\,\mathbf{s}\|_1\\ \text{s. t.} \quad & \|\hat{\mathbf{\Sigma}}\,\mathbf{s}\|_2 \leq \epsilon_n, \quad \mathbf{s}_{\mathcal{D}} = \mathbf{0}, \quad \mathbf{s}_{\mathcal{L}} = \mathbf{s}_{\mathcal{U}},\\ & (\mathbf{e}_1\otimes\mathbf{1}_N)^\top\mathbf{s} = 1,\end{aligned} \tag{13}$$

where the second, third, and fourth constraints correspond to the feasibility conditions in (11). Decomposing $\mathbf{s}$ into $\mathbf{s}_{\mathcal{D}}$, $\mathbf{s}_{\mathcal{L}}$, and $\mathbf{s}_{\mathcal{U}}$, we may write the first constraint in (13) as $\|[\hat{\mathbf{\Sigma}}]^\top_{\mathcal{D}}\,\mathbf{s}_{\mathcal{D}} + [\hat{\mathbf{\Sigma}}]^\top_{\mathcal{L}}\,\mathbf{s}_{\mathcal{L}} + [\hat{\mathbf{\Sigma}}]^\top_{\mathcal{U}}\,\mathbf{s}_{\mathcal{U}}\|_2 \leq \epsilon_n$. This enables us to restate (13) only in terms of $\mathbf{s}_{\mathcal{L}}$ as follows

$$\begin{aligned}\hat{\mathbf{s}}_{\mathcal{L}} \quad = \quad & \underset{\mathbf{s}_{\mathcal{L}}}{\mathrm{argmin}} \ \|\mathbf{R}\mathbf{s}_{\mathcal{L}}\|_1\\ \text{s. t.} \quad & \|\mathbf{M}\,\mathbf{s}_{\mathcal{L}}\|_2 \leq \epsilon_n, \quad (\mathbf{e}_1\otimes\mathbf{1}_{N-1})^\top\mathbf{s}_{\mathcal{L}} = 1,\end{aligned} \tag{14}$$

where we have assumed that N is even to simplify the notation in the last constraint in (14). We now introduce a slight variation on problem (14) parametrized by $q > 0$, where we relax

the equality constraint as follows

$$\begin{aligned} \hat{\mathbf{s}}_{\mathcal{L}}^{(q)} \;=\; & \underset{\mathbf{s}_{\mathcal{L}}}{\operatorname{argmin}} \;\; \|\mathbf{R}\mathbf{s}_{\mathcal{L}}\|_1 \\ & \text{s. t.} \qquad \left\| \begin{bmatrix} \mathbf{M} \\ q\,(\mathbf{e}_1 \otimes \mathbf{1}_{N-1})^\top \end{bmatrix} \mathbf{s}_{\mathcal{L}} - \begin{bmatrix} \mathbf{0} \\ q \end{bmatrix} \right\|_2 \leq \epsilon_n. \end{aligned} \tag{15}$$

Notice that parameter q controls the admissible level of violation of the original equality constraint in (14). In particular, for large q the equality must hold, i.e., $\lim_{q\to\infty} \hat{\mathbf{s}}_{\mathcal{L}}^{(q)} = \hat{\mathbf{s}}_{\mathcal{L}}$. For notational convenience, let us define $\mathbf{t}_q = q\,(\mathbf{e}_1 \otimes \mathbf{1}_{N-1})$, $\mathbf{\Phi}_q = [\mathbf{M}^\top, \mathbf{t}_q]^\top$, and $\mathbf{b}_q = [\mathbf{0}^\top, q]^\top$, where we explicitly state their dependence on the parameter q. In Claim 1 we prove recovery conditions for problem (15), where the parameter q plays a central role. More precisely, we bound the distance between the solution $\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}$ for (15) and the true graph $\mathbf{s}_{\mathcal{L}}^*$. The proof of this claim is deferred to the appendix.

Claim 1 *If the following two conditions are satisfied:*
l1) $\mathbf{\Phi}_q$ *is full column rank.*
l2) $\|\mathbf{\Phi}_q \mathbf{s}_{\mathcal{L}}^* - \mathbf{b}_q\|_2 \leq \epsilon_n$.
Then, we have that

$$\begin{aligned} & \|\hat{\mathbf{s}}_{\mathcal{L}}^{(q)} - \mathbf{s}_{\mathcal{L}}^*\|_1 \leq \gamma_q \epsilon_n, \\ & \textit{where } \gamma_q = \frac{2\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})\|\mathbf{R}^\dagger\|_1}{\sigma_{\min}(\mathbf{\Phi}_q)} \left(2 + \sqrt{|\mathcal{K}|}\right). \end{aligned} \tag{16}$$

We now show that requirements *1)-5)* in the statement of Theorem 2 imply conditions *l1)* and *l2)* in Claim 1 as $q \to \infty$. That *1)* implies *l1)* follows from a simple argument. Indeed, given that $\mathbf{\Phi}_q$ is generated from $\mathbf{M}$ by adding the row corresponding to $q\,(\mathbf{e}_1 \otimes \mathbf{1}_{N-1})^\top$, the column rank of $\mathbf{\Phi}_q$ cannot be smaller than that of $\mathbf{M}$. Since $\mathbf{M}$ is full column rank, $\mathbf{\Phi}_q$ must be as well. That *2)-5)* imply *l2)* is shown in the following claim, whose proof is also deferred to the appendix.

Claim 2 *If conditions* 2)-5) *in the statement of Theorem 2 hold, then with probability at least* $1 - e^{-C' \log N}$ *for some constant* $C' > 0$ *we have that* $\|\mathbf{\Phi}_q \mathbf{s}_{\mathcal{L}}^* - \mathbf{b}_q\|_2 \leq \epsilon_n$ *as* $q \to \infty$.

Recall that the solution $\hat{\mathbf{s}}_{\mathcal{L}}$ of problem (14) coincides with $\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}$ for $q \to \infty$. Hence, from Claims 1 and 2, it follows that under the conditions of Theorem 2 it holds with high probability that $\|\hat{\mathbf{s}}_{\mathcal{L}} - \mathbf{s}_{\mathcal{L}}^*\|_1 \leq \gamma_\infty \epsilon_n$, where $\gamma_\infty := \lim_{q\to\infty} \gamma_q$. Moreover, in terms of the full matrices (instead of just the lower triangular components), this implies that

$$\sum_{k=1}^{K} \|\mathrm{vec}(\hat{\mathbf{S}}^{(k)} - \mathbf{S}^{(k)^*})\|_1 \leq 2\,\gamma_\infty \epsilon_n. \tag{17}$$

Consequently, if we show that $2\,\gamma_\infty \leq \gamma$ as defined in (12) the proof concludes. More specifically, we want to show that

$$\lim_{q\to\infty} \frac{4\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})\|\mathbf{R}^\dagger\|_1}{\sigma_{\min}(\mathbf{\Phi}_q)} \left(2 + \sqrt{|\mathcal{K}|}\right) \leq \frac{4\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})\|\mathbf{R}^\dagger\|_1}{\sigma_{\min}(\mathbf{M})} \left(2 + \sqrt{|\mathcal{K}|}\right). \tag{18}$$

This boils down to showing that $\lim_{q\to\infty} \sigma_{\min}(\mathbf{\Phi}_q) \geq \sigma_{\min}(\mathbf{M})$. To prove this, we use Lemma 1. Let us denote the eigendecomposition of the real and symmetric matrix $\mathbf{M}^\top\mathbf{M} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$. Notice that the singular values of $\mathbf{M}$ are then given by $\sqrt{d_i}$. From the definition of $\mathbf{\Phi}_q$ it readily follows that $\mathbf{\Phi}_q^\top\mathbf{\Phi}_q = \mathbf{M}^\top\mathbf{M} + \mathbf{t}_q\mathbf{t}_q^\top$. We may rewrite this equality in a form more amenable to Lemma 1 as $\mathbf{V}^\top\mathbf{\Phi}_q^\top\mathbf{\Phi}_q\mathbf{V} = \mathbf{D} + \mathbf{V}^\top\mathbf{t}_q\mathbf{t}_q^\top\mathbf{V}$. Equating $\mathbf{V}^\top\mathbf{\Phi}_q^\top\mathbf{\Phi}_q\mathbf{V}$ to $\mathbf{C}$ and $\mathbf{V}^\top\mathbf{t}_q$ to $\mathbf{u}$ in Lemma 1, we obtain that $\sigma_{\min}(\mathbf{\Phi}_q) \geq \sigma_{\min}(\mathbf{M})$ for all $q > 0$. In particular, $\lim_{q\to\infty} \sigma_{\min}(\mathbf{\Phi}_q) \geq \sigma_{\min}(\mathbf{M})$ as we wanted to show, concluding the proof of the theorem. ■

Theorem 2 provides a high-probability bound on the error incurred when solving (11) for graph signals resulting from a Gaussian white noise process. The fact that the result is probabilistic in nature is expected. Indeed, since we are estimating the covariances from observed signals, there is always a small chance that the estimates are too noisy to enable approximate recovery of the GSOs. Let us now analyze the five conditions required in the statement of the theorem. Condition *1)* is akin to requiring the feasibility set to be a singleton when perfect covariances are available. To see this, notice that if in (14) we make $\epsilon_n = 0$ and $\mathbf{M}$ is full column rank, then $\mathbf{s}_{\mathcal{L}}$ is completely determined. Relating this back to Theorem 1, recovery in the noiseless case is guaranteed in this setting [cf. discussion after Theorem 1]. However, the current theorem describes how this recovery degrades with noise in the estimated covariances. Condition *3)* imposes the reasonable requirement that the amount of signals observed from each graph is comparable. Intuitively, since our objective is to gain inference power by pooling signals together, the case where only a vanishing number of signals are associated with a specific graph is detrimental to the estimation of that graph. Conditions *2)* and *4)* impose relations between the size of the graphs N, the number of signals available n, and the number of graphs K. The number of graphs should be small in relation to the number of nodes in each of those graphs [cf. condition *2)*] and the number of nodes cannot be too large compared with the number of observed signals [cf. condition *4)*]. Condition *5)* provides a direct handle on the recovery error by determining the minimum admissible ϵ_n. More precisely, if ϵ_n is too small then the problem might become infeasible or no approximate solution might be included in the feasibility set. On the other hand, if ϵ_n is chosen too large then the bound in (12) would be too loose. In this context, condition *5)* guides the choice of ϵ_n so that it is large enough for the result to hold while trying to minimize the upper bound on the estimation error. Consistent with our discussion of condition *1)*, whenever $n \to \infty$, we have that $\epsilon_n \to 0$ and (12) guarantees a perfect recovery. More interestingly, Theorem 2 reveals the behavior of this error for finite values of n. Indeed, for fixed K and N, ϵ_n decreases as $1/\sqrt{n}$ and the only term in γ dependent on n is $\sigma_{\min}(\mathbf{M})$. The revealed functional dependence arises in practice, as we illustrate in the next section.

5. Numerical Experiments

Through synthetic and real-world graphs, we validate our theoretical claims and illustrate the performance of the proposed method for joint inference of networks.

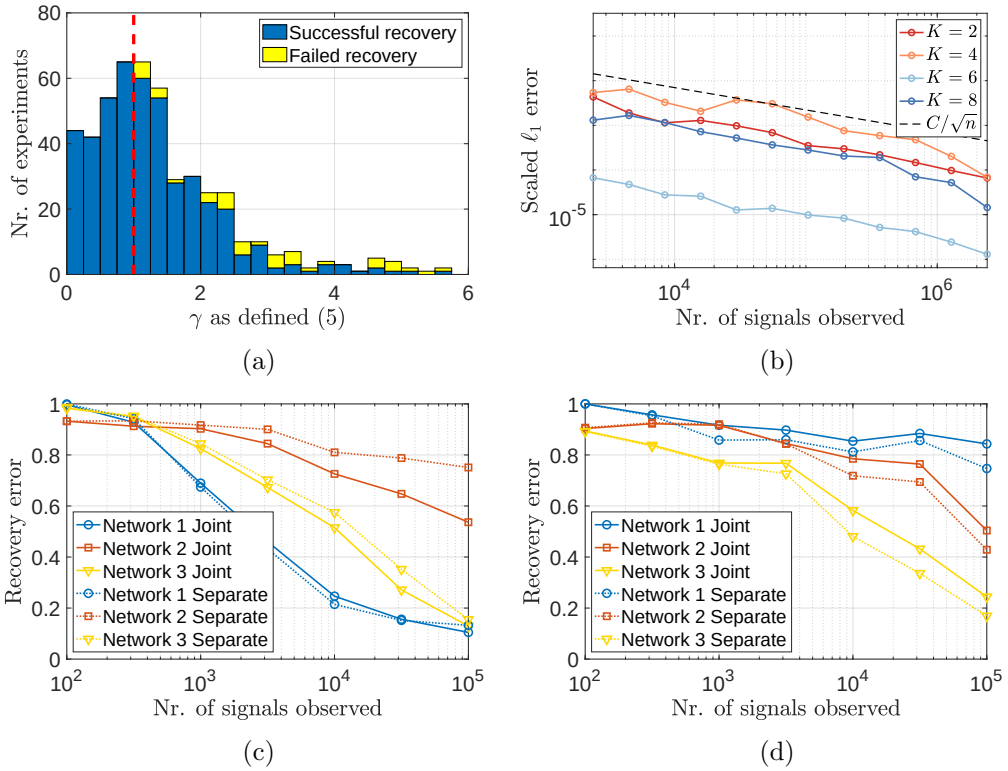


Figure 1: (a) Experimental validation of Theorem 1. For every realization where γ in (5) is strictly less than 1, perfect recovery is achieved. (b) Experimental validation of Theorem 2. The scaled sum of the ℓ_1-norm recovery errors decreases as $1/\sqrt{n}$ as larger numbers of signals are observed. (c) Recovery error for three social networks with similar structure as a function of the number of signals in the computation of the sample covariance. The joint inference method in (11) achieves lower overall error than the separate inference of each network. (d) Recovery error for three networks with no common structure as a function of the number of signals in the computation of the sample covariance. By enforcing a non-existent similarity between networks, the joint inference method underperforms compared to the separate inference.

5.1 Conditions for Noiseless Recovery

In this numerical experiment, we illustrate the theoretical guarantees in Theorem 1 by jointly inferring pairs of networks from perfect knowledge of the covariances of graph stationary processes. More specifically, we generate 500 pairs of graphs where one graph in each pair is generated from an Erdős-Rényi model (Bollobás 2001) of size $N = 20$ and edge-formation probability $p = 0.1$, and the other graph is obtained by randomly rewiring 3 edges of the first one. Notice that this procedure ensures that both graphs in each pair are similar, thus motivating our joint inference method. Covariance matrices of stationary processes in each graph are generated randomly by constructing filters $\mathbf{H}$ (see Section 1.5) of size $L = 3$ with normally distributed coefficients, and then setting $\mathbf{C_x} = \mathbf{H}\mathbf{H}^\top$. Our goal is to recover the adjacency matrices of each pair of graphs by solving 500 instances of (3),

where we set $\alpha_1 = \alpha_2 = \beta_{1,2} = 1$. For each of these 500 attempts we record whether the recovery was successful or not, whether condition 1 in Theorem 1 was satisfied or not—it was satisfied in all cases—, and the value of γ in (5). In Figure 1(a) we plot the histogram of γ discriminating by recovery performance. The figure clearly depicts the result of Theorem 1 in that, for all cases in which $\gamma < 1$, relaxation (3) achieves perfect recovery. Equally important, Figure 1(a) reveals that the bound stated in (5) is tight since some realizations with $\gamma = 1$ led to failed recoveries as indicated by the yellow portion of the bar to the right of the dashed line.

5.2 High-Probability Error Bound

We next demonstrate the upper bound for the recovery error in Theorem 2 in the case of fixed nodes N and graphs K. We generate one graph from an Erdős-Rényi model with $N = 20$ and edge-formation probability $p = 0.4$, and the remaining $K - 1$ graphs are obtained by rewiring the edges of the first graph with probability $q = 0.3$. Graph filters are constructed as in the previous experiment. In this case, we are demonstrating robust graph inference using sample covariance matrices obtained from an increasing number of observed graph signals. Since K and N are fixed, we would expect the ℓ_1-norm recovery errors to be dependent on $C/\sqrt{n}$ through the constant ϵ_n and $C'/\sigma_{\min}(\mathbf{M})$ through γ for proper choices of the constants C and C' [cf. (12)]. Scaling the sum of the ℓ_1-norm recovery errors by $\sigma_{\min}(\mathbf{M})$ would be expected to be upper bounded by $C/\sqrt{n}$, where C is the appropriately selected constant. This is observed in practice, as portrayed in Figure 1(b), where we plot the recovery error of $K \in \{2, 4, 6, 8\}$ graphs as the number of signals increases. The error shown in the figure corresponds to the sum in (12) normalized by the ℓ_1-norm sum of the true GSOs and scaled by $\sigma_{\min}(\mathbf{M})$, i.e., $\sigma_{\min}(\mathbf{M}) \sum_{k=1}^{K} \|\hat{\mathbf{S}}^{(k)} - \mathbf{S}^{(k)^*}\|_1 / \sum_{k=1}^{K} \|\mathbf{S}^{(k)^*}\|_1$. In both cases, the scaled ℓ_1-norm error exhibits the expected monotonic descent as $C/\sqrt{n}$, as illustrated by the dashed line in Figure 1(b).

5.3 Joint Inference of Social Networks

Consider three graphs defined on a common set of nodes representing 32 students from the University of Ljubljana in Slovenia. The networks encode different types of interactions among the students, and were built by asking each student to select a group of preferred college mates for a number of situations, e.g., to discuss a personal issue or to invite to a birthday party[2]. The considered graphs are unweighted and symmetric, and the edge between i and j exists if either student i picked j in the questionnaire or vice versa. Notice that the obtained networks are naturally similar to each other since the choices of friends across different situations do not vary greatly. We test the recovery performance of the robust formulation in (11) where the sample covariances are estimated from varying numbers of graph signals, and ϵ_n is chosen as small as possible while ensuring feasibility. The graph signals are synthetically generated following covariance matrices obtained from the GSOs as explained in the previous numerical experiment. Figure 1(c) portrays the joint recovery errors for the three networks as the number of observed signals varies, and compares them to the corresponding errors obtained from inferring the networks separately. The error of

2. Access to the data and additional details are available at `http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:pajek:students`

an estimator $\hat{\mathbf{S}}$ is quantified as $\|\mathbf{S}-\hat{\mathbf{S}}\|_F/\|\mathbf{S}\|_F$, where $\mathbf{S}$ denotes the true GSO. First notice that for an increasing number of observed signals we see a monotonous decrease in recovery error. For instance, when going from 10^3 to 10^4 observations the error when inferring Network 1 is (approximately) divided by three. This is expected since a larger number of observations entails a more reliable estimate of the covariance matrix. More interestingly, we see an overall positive effect of the *joint* inference compared to the corresponding separate inferences. This effect is more conspicuous for Network 2, for which the inference based exclusively on its sample covariance has proven to be more challenging.

Finally, we repeat the above experiment but for three synthetically generated networks that model a scenario where the students choose their college mates completely at random. In this way, the similarity across the networks to be inferred is lost. Consequently, imposing this similarity in the joint inference problem is actually detrimental to the recovery performance as depicted in Figure 1(d). Indeed, from the figure it can be seen that for the three networks and for almost any possible number of observed signals the separate inference outperforms the joint method.

5.4 Varying the Number of Graphs and the Sparsity Level

We demonstrate the superiority of our proposed joint inference method in comparison with separate network inference over different settings by varying: (i) the number of graphs and (ii) the sparsity of the graphs. In Figure 2(a) we compare joint and separate network inference as the number of graphs to be estimated increases for $K \in \{2, 3, \ldots, 8\}$. Networks are generated by sampling an Erdős-Rényi graph with a fixed edge probability $p = 0.3$, and the K networks are generated by rewiring edges from the first network with probability $q = 0.1$. The total number of graph signals n is preserved for all sets of estimated graphs, so that the number of observed signals per graph is n/K. Correspondingly, we observe that the recovery error increases for larger K.

In the second experiment, we compare joint and separate network inference as the underlying network sparsity varies. The networks to be estimated are generated similarly as in previous experiment; one network is a sampled Erdős-Rényi network, and the K synthetic graphs are rewired versions of the first graph, where three edges are rewired in all cases. To vary the sparsity of the sets of graphs, the edge probability is observed in a range $p \in \{0.1, 0.2, \ldots, 0.9\}$. Estimation is performed for both joint and separate inference with the true covariance matrices to remove dependencies on noisy signals.

Observe that for both presented problem settings in Figures 2(a) and 2(b), increasing the number of graphs and varying the level of edge sparsity results in consistent superiority of joint network inference over separate inference. This illustrates the generality of the benefits of considering the joint estimation formulation.

5.5 Model Selection

The formulation of the problem in (3) presents a general case with freedom of different choices of α_k and $\beta_{k,k'}$ respectively for each graph and pair of graphs. These free parameters can be used to incorporate prior knowledge of the different levels of sparsity or graph similarities. In the example of time-varying network estimation, an appropriate choice of $\beta_{k,k'}$ may be to encourage networks adjacent in time to be more similar. In the absence of

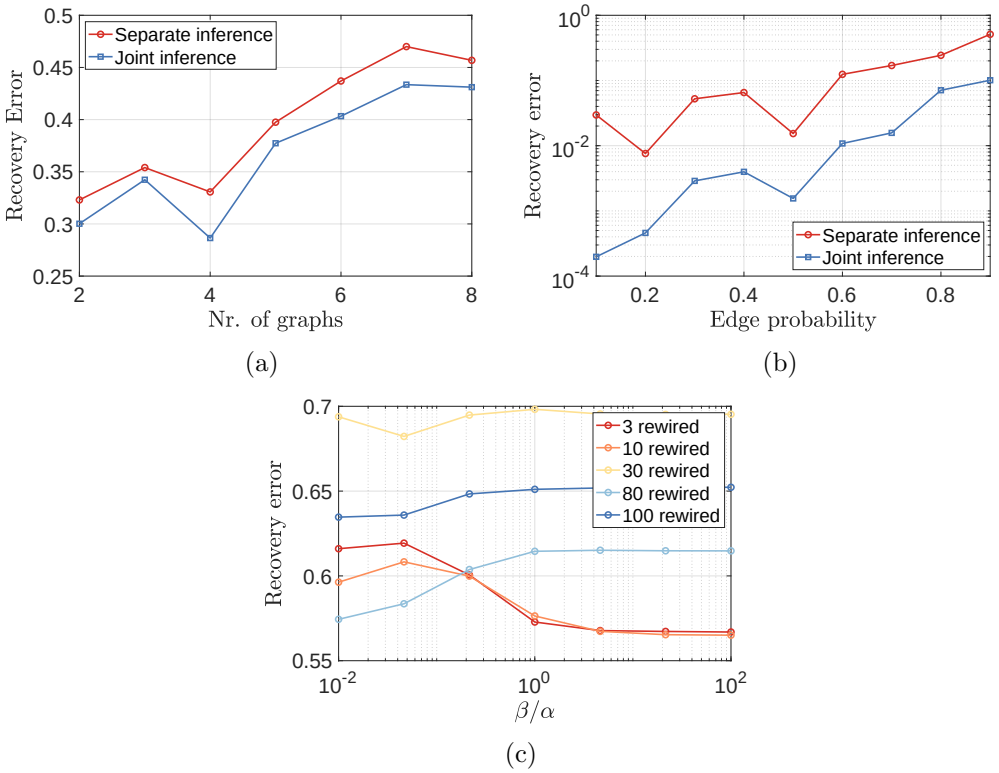
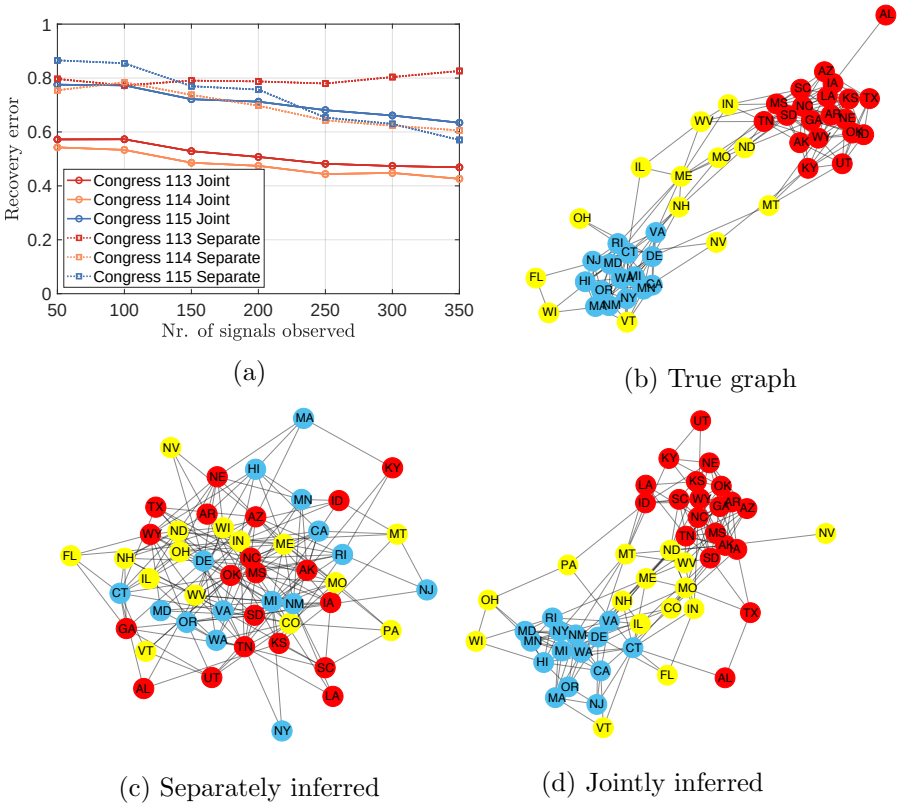


Figure 2: (a) Recovery error for multiple networks with similar structure as a function of the number of networks. The joint inference method consistently outperforms the separate inference method. (b) Recovery error for three similar networks as a function of the sparsity level of the networks. As the edge probability increases and the network sparsity decreases, the joint inference method continues to demonstrate a greater estimation performance. (c) Recovery error for three networks as a function of the tuning parameter determining the tradeoff between network sparsity and similarity. As the networks become more similar, increasing the parameter β/α to encourage pairwise network similarity results in greater estimation accuracy.

this prior knowledge, we implement the same value $\alpha_k = \alpha$ for all graphs k and the same value for $\beta_{k,k'} = \beta$ for all pairs of graphs (k, k'). In this case, note that there is only one degree of freedom determining the tradeoff between sparsity and similarity given by the ratio β/α. We now illustrate the effects of selecting different values of this ratio.

We jointly estimate $K = 3$ networks while observing effects on recovery performance as a function of β/α. The first graph is generated from an Erdős-Rényi model with $N = 25$ and edge-formation probability $p = 0.3$, and the remaining two graphs are obtained by rewiring a number of edges in the first graph. We emphasize the tradeoff between sparsity and similarity in our method by observing the varying parameter β/α as networks to be estimated become less similar. In particular, we present results for sets of networks that differ by an increasing number of edges, where we vary the number of edges that are rewired as $\{3, 10, 30, 80, 100\}$. When fewer edges are rewired, the true networks are more similar



(a) (b) True graph

(c) Separately inferred (d) Jointly inferred

Figure 3: (a) Mean recovery error over ten trials for three senate networks as a function of the number of signals considered in the computation of the sample covariance. Joint inference demonstrates less overall recovery error than separate inference of each senate network. (c) True graph of senate network for 114th congress with top-200 edges sorted by weight. Red nodes correspond to states with two Republican senators, blue nodes with two Democratic senators, and yellow nodes with senators from differing parties. (d) Separately recovered senate network for 114th congress with top-200 edges sorted by weight. Network recovery for a limited number of signals shows a mixed structure when each network is estimated alone. (e) Jointly recovered senate network for 114th congress with top-200 edges sorted by weight. A more similar structure to the true graph is observed when joint inference is applied for three similar senate networks.

thus a larger β/α leads to superior performance. The reverse also holds true, where graphs that differ by a greater number of edges cannot easily capitalize on the pairwise similarity penalty, thus a smaller β/α results in improved estimation accuracy.

5.6 Senate Networks

The comparison of joint and separate graph inference is also performed with real-world data of U.S. congress roll-call votes (Lewis et al. 2020). We observe the votes of 3 congresses, 113th (2013 to 2015), 114th (2015 to 2017), and 115th (2017 to 2019), from 2 senators per state (100 total). All $K = 3$ congresses are represented as networks, where senator opinions

per state are combined for $N = 50$ nodes shared by each graph. Nodal values for each state consist of the sum of the votes of both senators, where yea, nay, and other cases (such as abstention) are represented by 1, -1, and 0, respectively. The total number of roll-calls (graph signals) for the 113th, 114th, and 115th congresses are respectively 657, 502, and 599. Each state is separated into one of three categories based on the party affiliation of its senators. States are labeled as (i) D if both senators are in the Democratic Party, (ii) R if both senators are in the Republican Party, and (iii) M if senators are from different (mixed) parties.

In the absence of ground-truth senate networks, we deem as true underlying graphs those separately inferred for each congress when considering all the available graph signals. Moreover, to recover graphs on which the observed signals are not only stationary but also smooth, we add a regularization term $\|\mathbf{S} \circ \mathbf{Z}\|$ to the optimization objective, where $\mathbf{Z}$ is the pairwise distance matrix $Z_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$, and $\mathbf{x}_i$ contains the value of all signals at the i-th node; see Kalofolias (2016). Having established the ground-truth baselines, we perform joint and separate inference from limited observations and compare the estimation accuracy when gradually increasing the number of signals considered in the covariance computation. Subsets are randomly selected from all available votes, and ten trials of randomized subsets are performed to observe their mean behavior; see Figure 3(a).

Although the true networks were inferred separately, joint inference of senate networks markedly outperforms separate inference when a limited number of signals are available. This further reinforces our intuition that pooling observations from similar networks is especially relevant in data scarce settings. To better illustrate the difference in the inferred networks, in Figures 3(b) through 3(d) we provide spring layout plots of the true, separately inferred, and jointly inferred networks for congress 114th when 350 signals are observed. For clarity, only the top-200 edges sorted by weight are drawn. From the figures it becomes evident that the joint inference helps preserve the partisan structure of the true network whereas this important network feature is not recovered when performing separate inference.

6. Conclusions

We presented a method for jointly estimating multiple graphs from observed graph signals. The inference task was posited as a sparse recovery optimization problem regularized by the differences between the recovered graphs and subject to algebraic constraints derived from the assumption that the observed signals are stationary on the underlying graphs. A convex relaxation of the aforementioned optimization problem was presented and its tightness was shown under sufficient conditions for the case of perfect knowledge on the signal covariances. Furthermore, for the more relevant case where the covariances are estimated, a robust variation of the optimization problem is presented along with a high-probability bound on the recovery error. Finally, the results and intuition discussed throughout the paper were illustrated via numerical experiments on synthetic and real-world data.

Regarding potential avenues for future research, two generalizations of the setting here presented are of special interest. 1) It would be of interest to relax the assumption that we know on which graph each signal is defined. This case would require clustering the signals based on their estimated source graph and, most probably, an iterative formulation where the graphs are inferred and the signals reassigned between the graphs until convergence. 2)

It would be of interest to consider setups where the node sets (and their cardinality) are not the same across the graphs to be inferred. The rising popularity of graphons (Avella-Medina et al. 2020) may contribute to solving this setting, where the association of multiple graphs with a graphon presents a potential direction for joint graph inference with the underlying similarity between graphs dictated by the probability of being generated by a common graphon.

Acknowledgments

Research was supported by NSF (DMS-1651995 and CCF-2008555), and the Spanish Federal grants KLINILYCS (TEC2016-75361-R) and SPGraph (PID2019-105032GB-I00).

Appendix A. Proofs of Claims 1 and 2

We first state the following two lemmas that will be used to prove Claim 2.

Lemma 2 *(Cai et al., 2016, Lemma 2) Suppose* $\mathbf{r}_1, \cdots, \mathbf{r}_n$ *are* K*-dimensional random vectors satisfying* $\mathbb{E}[\mathbf{r}_i] = \mathbf{0}$ *and* $\|\mathbf{r}_i\|_2 \leq M$ *for* $1 \leq i \leq n$*. We have for any* $s > 0$ *and* $r > s$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \mathbf{r}_i\right\|_2 \geq r\right) \leq \mathbb{P}\left(\|\mathbf{z}\|_2 \geq (r-s)/\lambda_{\max}^{1/2}\right) + L, \tag{19}$$

where $L = c_1 K^{5/2} \exp(-c_2 K^{-5/2} s/M)$*,* $\lambda_{\max}$ *is the largest eigenvalue of* $Cov(\sum_{i=1}^{n} \mathbf{r}_i)$*,* $\mathbf{z}$ *is a* K*-dimensional standard normal random vector and* c_1, c_2 *are positive constants.*

Lemma 3 *Denoting by* $a^{(k)}$ *independent realizations of the random variable* $a \sim \mathcal{N}(0, \sigma^2)$*, the following tail bound holds*

$$\mathbb{P}\left(\frac{1}{m}\sum_{k=1}^{m}(a^{(k)})^2 - \mathbb{E}[a^2] \geq \sigma^2 t\right) \leq \exp\left(-\frac{m}{8}\min(t^2, t)\right). \tag{20}$$

Lemma 2 bounds in probability the sum of the norm of bounded random vectors whereas Lemma 3 is a standard result about tail bounds of chi-squared random variables. Having introduced these results, we can now show the two claims.

A.1 Proof of Claim 1

This proof has been partially inspired by Theorem 2 by Zhang et al. (2016). We are first going to show that condition *l1)* guarantees the existence of a vector $\mathbf{y} \in \mathbb{R}^{\left(K+\binom{K}{2}\right)N^2}$—that will be denominated *dual certificate*—such that $\mathbf{R}^\top \mathbf{y} \in \mathrm{Im}(\mathbf{\Phi}_q^\top)$, $\mathbf{y}_{\mathcal{K}} = \mathrm{sign}(\mathbf{R}_{\mathcal{K}} \mathbf{s}^*_{\mathcal{L}})$, and $\|\mathbf{y}_{\mathcal{K}^c}\|_\infty < 1$. In fact, we show here that we may attain that $\mathbf{y}_{\mathcal{K}^c} = \mathbf{0}$. Indeed, consider the vector $\mathbf{y}$ given by

$$\mathbf{y} = \mathbf{I}_{\mathcal{K}}^\top \mathrm{sign}(\mathbf{R}_{\mathcal{K}} \mathbf{s}^*_{\mathcal{L}}). \tag{21}$$

That $\mathbf{y}_{\mathcal{K}} = \mathrm{sign}(\mathbf{R}_{\mathcal{K}} \mathbf{s}^*_{\mathcal{L}})$, and $\|\mathbf{y}_{\mathcal{K}^c}\|_\infty = 0$ follow immediately from (21). Moreover, we have that $\mathbf{R}^\top \mathbf{y} \in \mathrm{Im}(\mathbf{\Phi}_q^\top)$ by realizing that $\mathbf{R}^\top \mathbf{y} = \mathbf{\Phi}_q^\top \mathbf{\Phi}_q (\mathbf{\Phi}_q^\top \mathbf{\Phi}_q)^{-1} \mathbf{R}^\top \mathbf{I}_{\mathcal{K}}^\top \mathrm{sign}(\mathbf{R}_{\mathcal{K}} \mathbf{s}^*_{\mathcal{L}})$,

where condition *l1)* guarantees the existence of the inverse. Now that we have established the existence of the dual certificate $\mathbf{y}$, it is helpful to notice that $\|\mathbf{R}\mathbf{s}^*_{\mathcal{L}}\|_1 = \mathbf{y}^\top \mathbf{R}\mathbf{s}^*_{\mathcal{L}}$.

We are ready to show the bound in (16). Consider an arbitrary vector $\mathbf{u} \in \mathbb{R}^{\binom{K+\binom{K}{2}}{N^2}}$ such that $\text{supp}(\mathbf{u}) \subseteq \mathcal{K}$. Letting $\boldsymbol{\rho} = \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}} - \mathbf{R}\mathbf{s}^*_{\mathcal{L}}$, $\boldsymbol{\rho}_1 = \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}} - \mathbf{u}$, and $\boldsymbol{\rho}_2 = \mathbf{R}\mathbf{s}^*_{\mathcal{L}} - \mathbf{u}$, we have that

$$\|\boldsymbol{\rho}\|_1 \leq \|\boldsymbol{\rho}_1\|_1 + \|\boldsymbol{\rho}_2\|_1. \tag{22}$$

We first focus on bounding the second summand in (22). By leveraging the fact that the support of $\boldsymbol{\rho}_2$ is contained in $\mathcal{K}$ we may write that

$$\begin{aligned} \|\boldsymbol{\rho}_2\|_1 \leq & \sqrt{|\mathcal{K}|}\ \|\boldsymbol{\rho}_2\|_2 \\ \leq & \sqrt{|\mathcal{K}|}\ \|\boldsymbol{\rho}\|_2 + \sqrt{|\mathcal{K}|}\ \|\boldsymbol{\rho}_1\|_2 \\ \leq & \sqrt{|\mathcal{K}|}\ \sigma_{\max}(\mathbf{R})\|\mathbf{s}^*_{\mathcal{L}} - \hat{\mathbf{s}}^{(q)}_{\mathcal{L}}\|_2 + \sqrt{|\mathcal{K}|}\ \|\boldsymbol{\rho}_1\|_1 \\ \leq & \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})}{\sigma_{\min}(\boldsymbol{\Phi}_q)}\ \|\boldsymbol{\Phi}_q(\mathbf{s}^*_{\mathcal{L}} - \hat{\mathbf{s}}^{(q)}_{\mathcal{L}})\|_2 + \sqrt{|\mathcal{K}|}\ \|\boldsymbol{\rho}_1\|_1, \end{aligned} \tag{23}$$

where in the third inequality we used that for an arbitrary vector $\mathbf{x}$ it holds that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$, and in the last inequality we used that $\|\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}\|_2/\sigma_{\min}(\mathbf{A})$, for every full column rank matrix $\mathbf{A}$. Condition *l1)* guarantees the validity of this operation.

We now find an upper bound for $\|\boldsymbol{\rho}_1\|_1$ for the vector $\mathbf{u}$ that minimizes this norm. More precisely, we want to bound $\xi := \min_{\mathbf{u}|\text{supp}(\mathbf{u})\subseteq\mathcal{K}} \|\boldsymbol{\rho}_1\|_1$. We may rewrite the support constraint on $\mathbf{u}$ as $\mathbf{I}_{\mathcal{K}^c}\mathbf{u} = \mathbf{0}$. Thus, the Lagrangian $L(\mathbf{u}, \mathbf{v})$ of the minimization problem becomes

$$\begin{aligned} L(\mathbf{u}, \mathbf{v}) &= \|\boldsymbol{\rho}_1\|_1 + \mathbf{v}^\top \mathbf{I}_{\mathcal{K}^c}\mathbf{u} \\ &= \|\boldsymbol{\rho}_1\|_1 + \mathbf{v}^\top \mathbf{I}_{\mathcal{K}^c}(\mathbf{u} - \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}) + \mathbf{v}^\top \mathbf{I}_{\mathcal{K}^c}\mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}. \end{aligned} \tag{24}$$

From duality theory we have that $\xi = \max_{\mathbf{v}} \min_{\mathbf{u}} L(\mathbf{u}, \mathbf{v})$. Moreover, if we define $\mathbf{w} := \mathbf{I}^\top_{\mathcal{K}^c}\mathbf{v}$, we have that

$$\xi = \max_{\mathbf{w}|\text{supp}(\mathbf{w})\subseteq\mathcal{K}^c} \min_{\mathbf{u}}\ \|\boldsymbol{\rho}_1\|_1 + \mathbf{w}^\top(\mathbf{u} - \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}) + \mathbf{w}^\top \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}. \tag{25}$$

By minimizing with respect to $\mathbf{u}$, for (25) not to result in $-\infty$, it must be that $\|\mathbf{w}\|_\infty \leq 1$. Otherwise, if $|w_r| > 1$ for some index r, the corresponding entry u_r can take a $-\infty$ value resulting in an unbounded minimization of ξ. In the case where $\|\mathbf{w}\|_\infty \leq 1$, the minimum for $\mathbf{u}$ is attained when $\mathbf{u} = \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}$. It thus follows that

$$\xi = \max_{\mathbf{w}|\,\text{supp}(\mathbf{w})\subseteq\mathcal{K}^c,\,\|\mathbf{w}\|_\infty\leq 1} \mathbf{w}^\top \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}. \tag{26}$$

Recalling that $\mathbf{y}$ is the previously introduced dual certificate [cf. (21)], we may write that

$$\xi = \max_{\mathbf{w}|\,\text{supp}(\mathbf{w})\in\mathcal{K}^c,\,\|\mathbf{w}\|_\infty\leq 1} (\mathbf{y} + \mathbf{w})^\top \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}} - \mathbf{y}^\top \mathbf{R}\hat{\mathbf{s}}^{(q)}_{\mathcal{L}}. \tag{27}$$

Moreover, since $\|\mathbf{y}\|_\infty = 1$, $\|\mathbf{w}\|_\infty \leq 1$ and the supports of $\mathbf{y}$ and $\mathbf{w}$ do not intersect, it readily follows that $\|\mathbf{y}+\mathbf{w}\|_\infty \leq 1$. Consequently, $(\mathbf{y}+\mathbf{w})^\top \mathbf{R}\hat{\mathbf{s}}_{\mathcal{L}}^{(q)} \leq \|\mathbf{R}\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}\|_1$. By substituting this in (27) we obtain that

$$\xi \leq \|\mathbf{R}\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}\|_1 - \mathbf{y}^\top \mathbf{R}\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}. \tag{28}$$

Leveraging the fact that $\|\mathbf{R}\mathbf{s}_{\mathcal{L}}^*\|_1 = \mathbf{y}^\top \mathbf{R}\mathbf{s}_{\mathcal{L}}^*$, we may write

$$\xi \leq \|\mathbf{R}\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}\|_1 - \|\mathbf{R}\mathbf{s}_{\mathcal{L}}^*\|_1 + \mathbf{y}^\top \mathbf{R}(\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)}) \leq \mathbf{y}^\top \mathbf{R}(\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)}), \tag{29}$$

where the inequality follows from $\|\mathbf{R}\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}\|_1 \leq \|\mathbf{R}\mathbf{s}_{\mathcal{L}}^*\|_1$ since $\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}$ is a minimizer of (15) whereas $\mathbf{s}_{\mathcal{L}}^*$ is a feasible solution of (15) due to condition *l2)*. Lastly, since we know that $\mathbf{R}^\top \mathbf{y}$ can be written as $\mathbf{R}^\top \mathbf{y} = \mathbf{\Phi}_q^\top \mathbf{\Phi}_q (\mathbf{\Phi}_q^\top \mathbf{\Phi}_q)^{-1} \mathbf{R}^\top \mathbf{I}_{\mathcal{K}}^\top \operatorname{sign}(\mathbf{R}_{\mathcal{K}} \mathbf{s}_{\mathcal{L}}^*)$, we may rewrite (29) as (recalling the definition of ξ)

$$\|\boldsymbol{\rho}_1\|_1 \leq \operatorname{sign}(\mathbf{R}_{\mathcal{K}}\mathbf{s}_{\mathcal{L}}^*)^\top \mathbf{I}_{\mathcal{K}} \mathbf{R}(\mathbf{\Phi}_q^\top \mathbf{\Phi}_q)^{-1} \mathbf{\Phi}_q^\top \mathbf{\Phi}_q (\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)}) \leq \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})}{\sigma_{\min}(\mathbf{\Phi}_q)} \|\mathbf{\Phi}_q(\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)})\|_2, \tag{30}$$

where the second inequality follows from the fact that every positive scalar is equal to its ℓ_2 norm. By substituting (30) back in (23) and then back in (22), we obtain that

$$\|\boldsymbol{\rho}\|_1 \leq \left(2 + \sqrt{|\mathcal{K}|}\right) \frac{\sqrt{|\mathcal{K}|}\sigma_{\max}(\mathbf{R})}{\sigma_{\min}(\mathbf{\Phi}_q)} \|\mathbf{\Phi}_q(\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)})\|_2. \tag{31}$$

Two observations are sufficient to obtain (16) from (31). First, notice that since $\mathbf{s}_{\mathcal{L}}^*$ and $\hat{\mathbf{s}}_{\mathcal{L}}^{(q)}$ both belong to the feasibility set of (15), we must have that $\|\mathbf{\Phi}_q(\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)})\|_2 \leq 2\epsilon_n$. Second, from compatibility of matrix induced norms and the fact that $\mathbf{R}$ is full column rank we have that $\|\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)}\|_1 = \|\mathbf{R}^\dagger \mathbf{R}(\mathbf{s}_{\mathcal{L}}^* - \hat{\mathbf{s}}_{\mathcal{L}}^{(q)})\|_1 \leq \|\mathbf{R}^\dagger\|_1 \|\boldsymbol{\rho}\|_1$.

A.2 Proof of Claim 2

Since $\sum_{j=1}^N {S_{j1}^{(1)}}^* = 1$, then $(\mathbf{e}_1 \otimes \mathbf{1}_N)^\top \mathbf{s}^* = 1$. Thus, having that $\|\mathbf{\Phi}_q \mathbf{s}_{\mathcal{L}}^* - \mathbf{b}_q\|_2 \leq \epsilon_n$ is equivalent to $\|\mathbf{M}\mathbf{s}_{\mathcal{L}}^*\|_2 \leq \epsilon_n$ for all q. From (11) this is equivalent to requiring that $\sum_{k=1}^K \|{\mathbf{S}^{(k)}}^* \hat{\mathbf{C}}^{(k)} - \hat{\mathbf{C}}^{(k)} {\mathbf{S}^{(k)}}^*\|_{\mathrm{F}}^2 \leq \epsilon_n^2$. Hence, we will show that this inequality holds with probability at least $1 - e^{-C' \log N}$ for some constant $C' > 0$ when conditions *2)-5)* in Theorem 2 hold.

Begin by noting that condition *4)* in particular implies that $\log N = o(n)$. This will be used throughout the proof. Leveraging the fact that ${\mathbf{S}^{(k)}}^* \mathbf{C}^{(k)} = \mathbf{C}^{(k)} {\mathbf{S}^{(k)}}^*$, and making use of the well-known inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we denote by $\mathbf{T}^{(k)} = {\mathbf{S}^{(k)}}^* \hat{\mathbf{C}}^{(k)} - \hat{\mathbf{C}}^{(k)} {\mathbf{S}^{(k)}}^*$, $\mathbf{T}_1^{(k)} = {\mathbf{S}^{(k)}}^* \hat{\mathbf{C}}^{(k)} - {\mathbf{S}^{(k)}}^* \mathbf{C}^{(k)}$, and $\mathbf{T}_2^{(k)} = \mathbf{C}^{(k)} {\mathbf{S}^{(k)}}^* - \hat{\mathbf{C}}^{(k)} {\mathbf{S}^{(k)}}^*$ so that

$$\left|[\mathbf{T}^{(k)}]_{ij}\right|^2 \leq 2\left|[\mathbf{T}_1^{(k)}]_{ij}\right|^2 + 2\left|[\mathbf{T}_2^{(k)}]_{ij}\right|^2$$

for all i, j. In terms of the Frobenius norm of interest, this implies that

$$\begin{aligned}\sum_{k=1}^{K} \|\mathbf{T}^{(k)}\|_{\mathrm{F}}^2 \leq\ & N^2 \max_{i,j} \sum_{k=1}^{K} \left| \left[\mathbf{T}^{(k)}\right]_{ij} \right|^2 \\ \leq\ & 2N^2 \max_{i,j} \sum_{k=1}^{K} \left| \left[\mathbf{T}_1^{(k)}\right]_{ij} \right|^2 + 2N^2 \max_{i,j} \sum_{k=1}^{K} \left| \left[\mathbf{T}_2^{(k)}\right]_{ij} \right|^2 .\end{aligned} \tag{32}$$

We now focus bounding $\max_{i,j} \sum_{k=1}^{K} \left| \left[\mathbf{T}_1^{(k)}\right]_{ij} \right|^2$. This is sufficient, since an analogous procedure can be followed to bound the second summand in (32). In order to bound the first summand in (32), we are going to show that the random event

$$A := \left\{ \sum_{k=1}^{K} \left| \left[\mathbf{T}_1^{(k)}\right]_{ij} \right|^2 \leq c_\epsilon^2 \omega^2 K \frac{\log N}{n}, \quad \text{for all } i, j \right\}$$

holds with high probability for some constant $c_\epsilon > 0$. Notice that we can regard event A as the intersection of events specific to the entries (i, j), and consider the events

$$A_{ij} := \left\{ \sum_{k=1}^{K} \left| \left[\mathbf{T}_1^{(k)}\right]_{ij} \right|^2 \leq c_\epsilon^2 \omega^2 K \frac{\log N}{n} \right\}.$$

Recall that $\mathbf{X}^{(k)} \in \mathbb{R}^{N \times n_k}$ contains the signals $\mathbf{x}_i^{(k)}$ as columns. We denote as $(\mathbf{z}_j^{(k)})^\top \in \mathbb{R}^{n_k}$ the j-th row of $\mathbf{X}^{(k)}$, i.e., the vector collecting the value in the j-th position of each of the graph signals associated with the k-th GSO. Moreover, for simplicity we will denote by $(\mathbf{s}_i^{(k)})^\top \in \mathbb{R}^N$ the i-th row of $\mathbf{S}^{(k)^*}$. From (1), we then have that

$$\left| \left[\mathbf{T}_1^{(k)}\right]_{ij} \right| = \frac{1}{n_k} \left| (\mathbf{s}_i^{(k)})^\top \mathbf{X}^{(k)} (\mathbf{z}_j^{(k)}) - \mathbb{E}[(\mathbf{s}_i^{(k)})^\top \mathbf{X}^{(k)} (\mathbf{z}_j^{(k)})] \right|, \tag{33}$$

where, under a slight abuse of notation, we are now considering $\mathbf{X}^{(k)}$ and $\mathbf{z}_j^{(k)}$ as random variables instead of specific realizations. Given that the columns of $\mathbf{X}^{(k)}$ are i.i.d., we have that $(\mathbf{y}_i^{(k)})^\top := (\mathbf{s}_i^{(k)})^\top \mathbf{X}^{(k)} \sim \mathcal{N}(\mathbf{0}, (\mathbf{s}_i^{(k)})^\top \mathbf{C}^{(k)} \mathbf{s}_i^{(k)} \mathbf{I})$ and, by definition, $(\mathbf{z}_j^{(k)})^\top \sim \mathcal{N}(\mathbf{0}, [\mathbf{C}^{(k)}]_{jj} \mathbf{I})$. It then follows that each term in the sum $(\mathbf{y}_i^{(k)})^\top (\mathbf{z}_j^{(k)}) = \sum_{t=1}^{n_k} (\mathbf{y}_i^{(k)})_t (\mathbf{z}_j^{(k)})_t$, is i.i.d. Leveraging this decomposition, we may write

$$\left| \left[\mathbf{T}_1^{(k)}\right]_{ij} \right| = \left| \frac{1}{n_k} \sum_{t=1}^{n_k} y_{i,t}^{(k)} z_{j,t}^{(k)} - \mathbb{E}[y_i^{(k)} z_j^{(k)}] \right|,$$

where we denote by $y_i^{(k)}$ a scalar random variable representing the elements of $\mathbf{y}_i^{(k)}$ (since they are all i.i.d.) and as $y_{i,t}^{(k)}$ a specific realization of this random variable. The same applies for $z_j^{(k)}$ with respect to $\mathbf{z}_j^{(k)}$. We denote random variables $w_{i+j}^{(k)} = y_i^{(k)} + z_j^{(k)}$ and

$w_{i-j}^{(k)} = y_i^{(k)} - z_j^{(k)}$. By subsequently applying the identity $4ab = (a+b)^2 - (a-b)^2$ and the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ we obtain that

$$\left| \left[\mathbf{T}_1^{(k)} \right]_{ij} \right|^2 \leq \frac{1}{8} \left| \frac{1}{n_k} \sum_{t=1}^{n_k} (w_{i+j,t}^{(k)})^2 - \mathbb{E}[(w_{i+j}^{(k)})^2] \right|^2 + \frac{1}{8} \left| \frac{1}{n_k} \sum_{t=1}^{n_k} (w_{i-j,t}^{(k)})^2 - \mathbb{E}[(w_{i-j}^{(k)})^2] \right|^2 . \quad (34)$$

Observe that, since both $y_i^{(k)}$ and $z_j^{(k)}$ are Gaussian random variables, we have that $w_{i+j}^{(k)}$ and $w_{i-j}^{(k)}$ are also Gaussian with variance at most $4\omega_k$. We define $\rho_k := n_k/n$ and, for fixed i, j, we define $u_t^{(k)} := (\rho_k^{1/2}/n_k)\big((w_{i+j,t}^{(k)})^2 - \mathbb{E}[(w_{i+j}^{(k)})^2]\big)$ if $t \leq n_k$ and $u_t^{(k)} = 0$ if $t > n_k$. Also, let $\mathbf{u}_t := (u_t^{(1)}, \cdots, u_t^{(K)})^\top$. By definition, we can then write

$$\sum_{k=1}^{K} \rho_k \left(\left| \frac{1}{n_k} \sum_{t=1}^{n_k} (w_{i+j,t}^{(k)})^2 - \mathbb{E}[(w_{i+j}^{(k)})^2] \right| \right)^2 = \left\| \sum_{t=1}^{n} \mathbf{u}_t \right\|_2^2 .$$

Consider now a new event A'_{ij} based on the newly introduced variable $\mathbf{u}_t$, namely

$$A'_{ij} := \left\{ \left\| \sum_{t=1}^{n} \mathbf{u}_t \right\|_2^2 \leq c_\epsilon'^2 \omega^2 \frac{\log N}{n} \right\},$$

for some constant c'_ϵ. We now briefly show that there exists a constant c'_ϵ such that the probability of A'_{ij} occurring is not larger than the probability of occurrence of A_{ij}. Indeed, from condition *3)* we know that there exists some constant c_w such that $\rho_k \geq c_w/K$. Hence, when A'_{ij} occurs, it is also satisfied that

$$\sum_{k=1}^{K} \left| \frac{1}{n_k} \sum_{t=1}^{n_k} (w_{i+j,t}^{(k)})^2 - \mathbb{E}[(w_{i+j}^{(k)})^2] \right|^2 \leq (c_\epsilon'^2/c_w)\omega^2 K \frac{\log N}{n}. \quad (35)$$

A similar analysis can be used to bound the above expression but for $w_{i-j}^{(k)}$ instead of $w_{i+j}^{(k)}$. Hence, we substitute these bounds in the expression obtained by summing (34) over all $k = 1, \ldots, K$, to see that A_{ij} also occurs, where the constant c_ϵ in A_{ij} depends on c'_ϵ and c_w. Consequently, if we show that $\mathbb{P}(A'_{ij}) \geq 1 - c'e^{-c \log N}$ for some constants $c > 2$ and c', it would then follow that $\mathbb{P}(A_{ij}) \geq 1 - c'e^{-c \log N}$. Moreover, a union bound over all (i, j) then guarantees the existence of a constant $C' > 0$ such that $\mathbb{P}(A) \geq 1 - e^{-C' \log N}$. It follows from the discussion after (32) that this would complete the proof.

Consequently, we are left to show that under conditions *2)-5)* we have that $\mathbb{P}(A'_{ij}) \geq 1 - c'e^{-c \log N}$ for some constants $c > 2$ and c'. The remainder of the proof of Claim 2 is devoted to proving this statement. We are going to prove this by showing that $\mathbb{P}(\neg A'_{ij}) \leq c'e^{-c \log N}$.

Notice that we cannot directly use Lemma 2 to bound $\| \sum_{t=1}^{n} \mathbf{u}_t \|_2^2$, since we would need $\|\mathbf{u}_t\|_2$ to be bounded by some constant M. We therefore split A'_{ij} into two subevents and estimate the bound for the probability of each of the two subevents. The basic intuition is that, if we are on the random event

$$E_a := \left\{ |u_t^{(k)}| \leq (n \log N)^{-1/2} K^{1/2-a} \, \omega, \quad \text{for all } t, k \right\},$$

then the ℓ_2 norm of $\mathbf{u}_t$ would always be smaller than

$$\|\mathbf{u}_t\|_2 \leq M_a := (n \log N)^{-1/2} K^{1-a}\, \omega, \tag{36}$$

where a is a free parameter that will be fixed later in the proof.

In particular, if we split the complement of A'_{ij} as

$$\mathbb{P}(\neg A'_{ij}) \leq \mathbb{P}(\neg A'_{ij}|E_a)\mathbb{P}(E_a) + \mathbb{P}(\neg E_a), \tag{37}$$

we can use Lemma 2 to bound $\mathbb{P}(\neg A'_{ij}|E_a)$ and then use Lemma 3 to bound $\mathbb{P}(\neg E_a)$. Let us introduce a new variable

$$\hat{u}_t^{(k)} := u_t^{(k)} I\left\{|u_t^{(k)}| \leq (n \log N)^{-1/2} K^{1/2-a}\omega\right\} - \mathbb{E}\left[u_t^{(k)} I\left\{|u_t^{(k)}| \leq (n \log N)^{-1/2} K^{1/2-a}\omega\right\}\right]$$

and $\hat{\mathbf{u}}_t := (\hat{u}_t^{(1)}, \cdots, \hat{u}_t^{(K)})^\top$.

Notice that if we are on the random event E_a, then $\mathbf{u}_t$ and $\hat{\mathbf{u}}_t$ follow the same distribution except for a shift $v_t^{(k)} := \mathbb{E}\left[u_t^{(k)} I\left\{|u_t^{(k)}| \leq (n \log N)^{-1/2} K^{1/2-a}\omega\right\}\right]$. Putting it differently, the distribution of $u_t^{(k)} - \hat{u}_t^{(k)}$ is a constant $v_t^{(k)}$ when we are on the random event E_a.

We can use Lemma 3 to estimate the scale of $v_t^{(k)}$ with respect to N and n. To do this, first notice that $u_t^{(k)}$ is a chi-squared random variable with one degree of freedom. Thus, for some constant η we can apply Lemma 3 for $\sigma^2 = \frac{\sqrt{K}\omega}{8\eta n}$, $t = \frac{8\eta n l}{\sqrt{K}\omega}$ and $m = 1$, to obtain the tail bound

$$\mathbb{P}(u_t^{(k)} \geq l) \leq \exp\left(-\eta \frac{nl}{\sqrt{K}\omega}\right) \quad \text{with} \quad l \gg \frac{\sqrt{K}\omega}{n}. \tag{38}$$

Moreover, since $u_t^{(k)}$ has mean zero for all t and k, we have that

$$\begin{aligned} |v_t^{(k)}| &= \left|\mathbb{E}\left[u_t^{(k)} I\left\{|u_t^{(k)}| \leq (n \log N)^{-1/2} K^{1/2-a}\omega\right\}\right]\right| \\ &= \left|\mathbb{E}\left[u_t^{(k)} I\left\{|u_t^{(k)}| \geq (n \log N)^{-1/2} K^{1/2-a}\omega\right\}\right]\right|. \end{aligned} \tag{39}$$

It follows from the definition of $u_t^{(k)}$ that $u_t^{(k)} \geq -\rho_k^{1/2}(4\omega/n_k)$. From the fact that $\log N = o(n)$, we have that

$$\rho_k^{1/2} \frac{4\omega}{n_k} << (n \log N)^{-1/2} K^{1/2-a}\omega. \tag{40}$$

Then, combining both previous facts, when $u_t^{(k)}$ satisfies that $|u_t^{(k)}| \geq (n \log N)^{-1/2} K^{1/2-a}\omega$, it must be that $u_t^{(k)}$ is positive. Therefore, the right hand side of (39) can be further rewritten as $\mathbb{E}\left[u_t^{(k)} I\left\{u_t^{(k)} \geq (n \log N)^{-1/2} K^{1/2-a}\omega\right\}\right]$, where we have deleted the absolute value of $u_t^{(k)}$. We let $\theta := (n \log N)^{-1/2} K^{1/2-a}\omega$ and $\gamma := \eta \frac{n}{\sqrt{K}\omega}$. Consequently, we may bound $|v_t^{(k)}|$ as follows

$$\begin{aligned} |v_t^{(k)}| &= \mathbb{E}\left[u_t^{(k)} I\left\{u_t^{(k)} \geq (n \log N)^{-1/2} K^{1/2-a}\omega\right\}\right] \\ &\leq \int_\theta^\infty \exp(-\gamma \ell)\, d\ell = \frac{1}{\gamma} \exp(-\gamma\theta) \\ &= \frac{\sqrt{K}\omega}{\eta n} \exp\left(-\eta (n/\log N)^{1/2} K^{-a}\right), \end{aligned}$$

where we have used (38) in the computation of the expected value. Clearly, $|v_t^{(k)}|$ decays exponentially with respect to $n/\log N$. Therefore, we have that $n|v_t^{(k)}| = o(\sqrt{(\log N)/n})$ for all $k = 1, \dots, K$. In addition, when we are on the random event E_a, we have that $\sum_{t=1}^n u_t^{(k)} = \sum_{t=1}^n \hat{u}_t^{(k)} + n_k v_t^{(k)}$ and therefore $(\sum_{t=1}^n u_t^{(k)})^2/2 \leq (\sum_{t=1}^n \hat{u}_t^{(k)})^2 + (n_k v_t^{(k)})^2$. By summing the previous expression over all $k = 1, \dots, K$, we further have that

$$\frac{1}{2}\left\|\sum_{t=1}^n \mathbf{u}_t\right\|_2^2 \leq \sum_{k=1}^K (n_k\, v_t^{(k)})^2 + \left\|\sum_{t=1}^n \hat{\mathbf{u}}_t\right\|_2^2. \tag{41}$$

By combining (41) with the fact that $n|v_t^{(k)}| = o(\sqrt{(\log N)/n})$ for all k, we further have that there exists some $0 < \delta < 1$ such that if we are on the event E_a, then $\|\sum_{t=1}^n \mathbf{u}_t\|_2 \geq c'_\epsilon \omega \sqrt{\frac{\log N}{n}}$ indicates that $\|\sum_{t=1}^n \hat{\mathbf{u}}_t\|_2 \geq (1-\delta) c'_\epsilon \omega \sqrt{\frac{\log N}{n}}$. Equivalently, there exists some constant $0 < \delta < 1$ such that, given the event,

$$B = \left\{ \left\| \sum_{t=1}^n \hat{\mathbf{u}}_t \right\|_2 \geq (1-\delta) c'_\epsilon \omega \sqrt{\frac{\log N}{n}} \right\},$$

the following inequality holds

$$\mathbb{P}(\neg A'_{ij} | E_a)\mathbb{P}(E_a) \leq \mathbb{P}(B \mid E_a)\mathbb{P}(E_a) \leq \mathbb{P}(B), \tag{42}$$

where the second inequality follows readily from Bayes' theorem. Given that $\hat{\mathbf{u}}_t$ is obtained from $\mathbf{u}_t$ by cutting the tails, we have that $\lambda_{\max}\{\mathrm{Cov}(\sum_{t=1}^n \hat{\mathbf{u}}_t)\} \leq \lambda_{\max}\{\mathrm{Cov}(\sum_{t=1}^n \mathbf{u}_t)\}$. In addition, as $u_t^{(k)}$ are i.i.d. for all t, we have that

$$\lambda_{\max}\left\{\mathrm{Cov}\left(\sum_{t=1}^n \hat{\mathbf{u}}_t\right)\right\} = \max_k \mathrm{Var}\left(\sum_{t=1}^n u_t^{(k)}\right) = (n_k \rho_k / n_k^2)\mathrm{Var}\left((w_{i+j}^{(k)})^2\right) \leq 16\omega^2/n.$$

We may thus apply Lemma 2 to further bound (42) by setting $r = (1-\delta)c'_\epsilon \omega \sqrt{\frac{\log N}{n}}$, $s = \frac{1}{2}(1-\delta)c'_\epsilon \omega \sqrt{\frac{\log N}{n}}$, $\lambda_{\max} = 16\omega^2/n$ and $M = M_a$ as defined in (36) to get that

$$\mathbb{P}(B) \leq \quad \mathbb{P}(\|\mathbf{z}\|_2 \geq C_2\sqrt{\log N}) + C_3 \exp\left\{\frac{2}{5}\log K - C_4 K^{a-7/2}\log N\right\}, \tag{43}$$

where C_2 and C_4 are constants that increase with increasing c_ϵ and $\mathbf{z}$ is a K-dimensional standard normal random vector. Note that the constant δ is also absorbed into C_2 and C_4. From condition *2)* and the tail bound in Lemma 3 we get that

$$\mathbb{P}(\|\mathbf{z}\|_2 \geq C_2\sqrt{\log N}) \leq C_1 \exp(-C'_2(\log N - K)),$$

where C'_2 is a constant that increases with increasing of c_ϵ. Replacing the above expression into (43) we obtain

$$\mathbb{P}(B) \leq \exp\{-C'_2 \log N\} + C_3 \exp\left\{\frac{2\log K}{5} - \frac{C_4 \log N}{K^{7/2-a}}\right\},$$

where C_2' and C_4 are the constants that would increase with the increment of c_ϵ. In this case, by choosing $a = 7/2$, with constant c_ϵ big enough, there exists some constant $c_1 > 2$ such that

$$\mathbb{P}(B) \leq \exp(-c_1 \log N). \tag{44}$$

Replacing (44) into (42) gives us the sought exponential bound for the first summand in (37). We are now left with the task of finding a bound for $\mathbb{P}(\neg E_a)$.

Given event $B'_{(k)} = \left\{ |u_t^{(k)}| \geq \left(\frac{K^{1-2a}}{n \log N}\right)^{1/2} \omega \right\}$, from the definition of the event E_a it follows that

$$\mathbb{P}(\neg E_a) \leq K \left(\max_{1 \leq k \leq K} n_k \right) \max_{1 \leq k \leq K} \mathbb{P}\left(B'_{(k)}\right). \tag{45}$$

From the relationship in (40) we obtain that the probabilities $\mathbb{P}\left(u_t^{(k)} \geq \left(\frac{K^{1-2a}}{n \log N}\right)^{1/2} \omega\right)$ and $\mathbb{P}\left(|u_t^{(k)}| \geq \left(\frac{K^{1-2a}}{n \log N}\right)^{1/2} \omega\right)$ are the same. Plus, from the tail bound in (38) we get

$$\mathbb{P}\left(B'_{(k)}\right) \leq e^{-\eta \sqrt{\frac{n}{K^{2a} \log N}}} = e^{-\eta \sqrt{\frac{n}{K^7 \log N}}},$$

where the last equality follows from recalling that we have fixed $a = 7/2$ in order to write (44). Condition *3)* guarantees the existence of some constant c_w such that $n_k \leq n c_w / K$ for all k, thus

$$\mathbb{P}(\neg E_a) \leq c_w n e^{-\eta \sqrt{\frac{n}{K^7 \log N}}} \leq c_w e^{\log n - \eta \sqrt{\frac{n}{K^7 \log N}}}. \tag{46}$$

From condition *4)* it follows that $\log n = o(\sqrt{n/(K^7 \log N)})$ and $\log N = o(\sqrt{n/(K^7 \log N)})$, which immediately implies that $\log N = o\left(\sqrt{\frac{n}{K^7 \log N}} - \log n\right)$. Combining this expression with (46) reveals that

$$\mathbb{P}(\neg E_a) \leq c_w e^{-c_2 \log N} \tag{47}$$

for some constant $c_2 > 2$, thus obtaining an exponential bound for the second summand in (37). To conclude, from the combination of (44) and (47) we get that $\mathbb{P}(\neg A'_{ij}) \leq c' \exp(-c \log N)$ for some $c > 2$, as wanted.

References

Hesam Araghi, Mohammad Sabbaqi, and Massoud Babaie-Zadeh. K-graphs: An algorithm for graph signal clustering and multiple graph learning. *IEEE Signal Process. Lett.*, 26 (10):1486–1490, 2019.

Jesús Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E. Priebe, and Joshua T. Vogelstein. Inference for multiple heterogeneous networks with a common invariant subspace. *J. Mach. Learn. Res. (JMLR)*, 22(142):1–49, 2021.

Marco Avella-Medina, Francesca Parise, Michael T. Schaub, and Santiago Segarra. Centrality measures for graphons: Accounting for uncertainty in networks. *IEEE Trans. Network Science and Eng.*, 7(1):520–537, 2020.

Brian Baingana and Georgios B. Giannakis. Tracking switched dynamic network topologies from information cascades. *IEEE Trans. Signal Process.*, 65(4):985–997, Feb 2017. ISSN 1053-587X. doi: 10.1109/TSP.2016.2628354.

Brian Baingana, Gonzalo Mateos, and Georgios B. Giannakis. Proximal-gradient algorithms for tracking cascades over social networks. *IEEE J. Sel. Topics Signal Process.*, 8:563–575, August 2014.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res. (JMLR)*, 9(Mar):485–516, 2008.

Anders E. Bilgrau, Carel F. W. Peeters, Poul Svante Eriksen, Martin Bøgsted, and Wessel N. van Wieringen. Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *J. Mach. Learn. Res. (JMLR)*, 21(26):1–52, 2020.

P. V. Bindu, P. Santhi Thilagam, and Deepesh Ahuja. Discovering suspicious behavior in multilayer social networks. *Computers in Human Behavior*, 73:568–582, 2017.

Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001.

T. Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445, 2016.

Xiaodong Cai, Juan Andrés Bazerque, and Georgios B. Giannakis. Sparse structural equation modeling for inference of gene regulatory networks exploiting genetic perturbations. *PLoS, Computational Biology*, June 2013.

Julien Chiquet, Yves Grandvalet, and Christophe Ambroise. Inferring multiple graphical structures. *Stat. and Computing*, 21(4):537–553, 2011.

Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. Royal Stat. Soc: Series B (Stat. Methodol.)*, 76(2):373–397, 2014.

Petar Djuric and Cédric Richard. *Cooperative and Graph Signal Processing: Principles and Applications*. Academic Press, 2018.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning Laplacian matrix in smooth graph signal representations. *IEEE Trans. Signal Process.*, 64(23):6160–6173, 2016.

Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Process. Mag.*, 36(3):44–63, 2019.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Lingrui Gan, Xinming Yang, Naveen Narisetty, and Feng Liang. Bayesian joint estimation of multiple graphical models. In *Advances in Neural Info. Process. Syst.*, pages 9802–9812, 2019.

Benjamin Girault, Paulo Gonçalves, and Éric Fleury. Translation on graphs: An isometric shift operator. *IEEE Signal Process. Lett.*, 22(12):2416–2420, 2015.

Gene H. Golub. Some modified matrix eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

Min Jin Ha, Francesco Claudio Stingo, and Veerabhadran Baladandayuthapani. Bayesian structure learning in multilayered genomic networks. *J. American Stat. Assoc.*, 116(534): 605–618, 2021.

Botao Hao, Will Wei Sun, Yufeng Liu, and Guang Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *J. Mach. Learn. Res. (JMLR)*, 18(1): 7981–8038, 2017.

Jian Hong and Xuchu Dai. Demixing and topology identification for mixed graph signals. In *IEEE Intl. Conf. Commun. and Info. Syst. (ICCIS)*, pages 154–159. IEEE, 2021.

Jean Honorio and Dimitris Samaras. Multi-task learning of Gaussian graphical models. In *Intl. Conf. on Machine Learning (ICML)*, pages 447–454, 2010.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Intl. Conf. Artif. Intel. Stat. (AISTATS)*, pages 920–929, 2016.

Vassilis Kalofolias, Andreas Loukas, Dorina Thanou, and Pascal Frossard. Learning time varying graphs. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 2826–2830, March 2017.

Jiun-Yu Kao, Dong Tian, Hassan Mansour, Antonio Ortega, and Anthony Vetro. Disc-GLasso: Discriminative graph learning with sparsity regularization. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 2956–2960. IEEE, 2017.

Georgios V. Karanikolas, Georgios B. Giannakis, Konstantinos Slavakis, and Richard M. Leahy. Multi-kernel based nonlinear models for connectivity identification of brain networks. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Shanghai, China, Mar. 20-25, 2016.

Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models.* Springer, New York, NY, 2009.

Sandeep Kumar, Jiaxi Ying, José Vinícius de Miranda Cardoso, and Daniel P. Palomar. A unified framework for structured graph learning via spectral constraints. *J. Mach. Learn. Res. (JMLR)*, 21(22):1–60, 2020.

Brenden M. Lake and Joshua B. Tenenbaum. Discovering structure by learning sparse graphs. In *Annual Cognitive Sc. Conf.*, pages 778 – 783, 2010.

Steffen L. Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

Wonyul Lee and Yufeng Liu. Joint estimation of multiple precision matrices with common structures. *J. Mach. Learn. Res. (JMLR)*, 16(1):1035–1062, 2015.

Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. Voteview: Congressional roll-call votes database. `https://voteview.com/`, 2020.

Anani Lotsi and Ernst Wit. Sparse Gaussian graphical mixture model. *Afrika Statistika*, 11(2):1041–1059, 2016.

Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. *J. Mach. Learn. Res. (JMLR)*, 17(1):5777–5824, 2016.

Hermina Petric Maretic and Pascal Frossard. Graph Laplacian mixture model. *IEEE Trans. Signal and Info. Process. over Networks*, 6:261–270, 2020.

Antonio G. Marques, Santiago Segarra, Geert Leus, and Alejandro Ribeiro. Stationary graph processes and spectral estimation. *IEEE Trans. Signal Process.*, 65(22):5911–5926, 2017.

Antonio G. Marques, Negar Kiyavash, José M. F. Moura, Dimitri Van De Ville, and Rebecca Willett. Graph signal processing: Foundations and emerging directions (editorial). *IEEE Signal Process. Mag.*, 37, Nov. 2020.

Gonzalo Mateos, Santiago Segarra, Antonio G. Marques, and Alejandro Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Process. Mag.*, 36(3):16–43, May 2019.

Jonathan Mei and José M. F. Moura. Signal processing on graphs: Estimating the structure of a graph. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 5495–5499, 2015.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.*, 34:1436–1462, 2006.

Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple Gaussian graphical models. *J. Mach. Learn. Res. (JMLR)*, 15(1): 445–488, 2014.

Yohsuke Murase, János Török, Hang-Hyun Jo, Kimmo Kaski, and János Kertész. Multilayer weighted social network model. *Physical Review E*, 90(5):052810, 2014.

Alberto Natali, Mario Coutino, Elvin Isufi, and Geert Leus. Online time-varying topology identification via prediction-correction algorithms. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 5400–5404. IEEE, 2021.

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proc. IEEE*, 106(5):808–828, 2018.

Brandon Oselio, Alex Kulesza, and Alfred O. Hero. Multi-layer graph analysis for dynamic social networks. *IEEE J. Sel. Topics Signal Process.*, 8(4):514–523, Aug. 2014. ISSN 1932-4553. doi: 10.1109/JSTSP.2014.2328312.

Bastien Pasdeloup, Vincent Gripon, Grégoire Mercier, Dominique Pastor, and Michael G. Rabbat. Characterization and inference of graph diffusion processes from observations of stationary signals. *IEEE Trans. Signal and Info. Process. over Networks*, 4(3):481–496, 2017.

Eduardo Pavez and Antonio Ortega. Generalized Laplacian precision matrix estimation for graph signal processing. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Shanghai, China, Mar. 20-25, 2016.

Carel F. W. Peeters, Anders Ellern Bilgrau, and Wessel N. van Wieringen. rags2ridges: A one-stop-shop for graphical modeling of high-dimensional precision matrices. *arXiv preprint arXiv:2010.05619*, 2020.

Nathanaël Perraudin and Pierre Vandergheynst. Stationary signal processing on graphs. *IEEE Trans. Signal Process.*, 65(13):3462–3477, 2017.

Christine Peterson, Francesco C. Stingo, and Marina Vannucci. Bayesian inference of multiple Gaussian graphical models. *J. American Stat. Assoc.*, 110(509):159–174, 2015.

Bradley S. Price, Aaron J. Molstad, and Ben Sherwood. Estimating multiple precision matrices with cluster fusion regularization. *Journal of Computat. and Graphical Stat.*, pages 1–12, 2021.

Ilaria Ricchi, Anjali Tarun, Hermina Petric Maretic, Pascal Frossard, and Dimitri Van De Ville. Dynamics of functional network organization through graph mixture learning. *bioRxiv*, 2021.

Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, and Vinod Menon. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861, 2012.

Aliaksei Sandryhaila and José M. F. Moura. Discrete signal processing on graphs. *IEEE Trans. Signal Process.*, 61(7):1644–1656, Apr. 2013. ISSN 1053-587X. doi: 10.1109/TSP.2013.2238935.

Stefania Sardellitti, Sergio Barbarossa, and Paolo Di Lorenzo. Enabling prediction via multilayer graph inference and sampling. In *Intl. Conf. on Sampling Theory and Applications (SampTA)*, pages 1–4. IEEE, 2019.

Stefania Sardellitti, Sergio Barbarossa, and Paolo Di Lorenzo. Online learning of time-varying signals and graphs. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 5230–5234. IEEE, 2021.

Santiago Segarra, Antonio G. Marques, Gonzalo Mateos, and Alejandro Ribeiro. Network topology inference from spectral templates. *IEEE Trans. Signal and Info. Process. over Networks*, 3(3):467–483, 2017a.

Santiago Segarra, Antonio G. Marques, and Alejandro Ribeiro. Optimal graph-filter design and applications to distributed linear network operators. *IEEE Trans. Signal Process.*, 65(15):4117–4131, 2017b.

Santiago Segarra, Gonzalo Mateos, Antonio G. Marques, and Alejandro Ribeiro. Blind identification of graph filters. *IEEE Trans. Signal Process.*, 65(5):1146–1159, 2017c.

Santiago Segarra, Yuhao Wang, Caroline Uhler, and Antonio G. Marques. Joint inference of networks from stationary graph signals. In *Asilomar Conf. on Signals, Systems, and Computers*, pages 975–979, 2017. doi: 10.1109/ACSSC.2017.8335493.

Yanning Shen, Brian Baingana, and Georgios B. Giannakis. Kernel-based structural equation models for topology identification of directed networks. *IEEE Trans. Signal Process.*, 65(10):2503–2516, 2017.

Olaf Sporns. *Discovering the Human Connectome*. MIT Press, Boston, MA, 2012.

Qinghua Tao, Xiaolin Huang, Shuning Wang, Xiangming Xi, and Li Li. Multiple Gaussian graphical estimation with jointly sparse penalty. *Signal Processing*, 128:88–97, 2016.

Katherine Tsai, Oluwasanmi Koyejo, and Mladen Kolar. Joint Gaussian graphical model estimation: A survey. *arXiv preprint arXiv:2110.10281*, 2021.

Gaël Varoquaux, Alexandre Gramfort, Jean-Baptiste Poline, and Bertrand Thirion. Brain covariance selection: Better individual functional connectivity models using population prior. In *Advances in Neural Info. Process. Syst.*, pages 2334–2342, 2010.

Yuhao Wang, Santiago Segarra, and Caroline Uhler. High-dimensional joint estimation of multiple directed Gaussian graphical models. *Electron. J. Statist.*, 14(1):2439–2483, 2020.

Koki Yamada and Yuichi Tanaka. Temporal multiresolution graph learning. *IEEE Access*, 9:143734–143745, 2021.

Koki Yamada, Yuichi Tanaka, and Antonio Ortega. Time-varying graph learning based on sparseness of temporal variation. In *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pages 5411–5415. IEEE, 2019.

Sen Yang, Zhaosong Lu, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2):916–943, 2015.

Xinming Yang, Lingrui Gan, Naveen N. Narisetty, and Feng Liang. GemBag: Group estimation of multiple Bayesian graphical models. *J. Mach. Learn. Res. (JMLR)*, 22(54): 1–48, 2021.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Hui Zhang, Ming Yan, and Wotao Yin. One condition for solution uniqueness and robustness of both l1-synthesis and l1-analysis minimizations. *Advances in Computat. Math.*, 42(6): 1381–1399, 2016.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.

Yu Zhu, Fernando J. Iglesias Garcia, Antonio G. Marques, and Santiago Segarra. Estimating network processes via blind identification of multiple graph filters. *IEEE Trans. Signal Process.*, 68:3049–3063, 2020a.

Yu Zhu, Michael T. Schaub, Ali Jadbabaie, and Santiago Segarra. Network inference from consensus dynamics with unknown parameters. *IEEE Trans. Signal and Info. Process. over Networks*, 6:300–315, 2020b.

Yunzhang Zhu and Lexin Li. Multiple matrix Gaussian graphs estimation. *J. Royal Stat. Soc: Series B (Stat. Methodol.)*, 80(5):927, 2018.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *J. American Stat. Assoc.*, 109(508):1683–1696, 2014.