

# Algorithms for Learning Graphs in Financial Markets

**José Vinícius de Miranda Cardoso**

*Department of Electronic and Computer Engineering  
The Hong Kong University of Science and Technology  
Clear Water Bay, Hong Kong*

JVDMC@CONNECT.UST.HK

**Jiaxi Ying**

*Department of Electronic and Computer Engineering  
The Hong Kong University of Science and Technology  
Clear Water Bay, Hong Kong*

JX.YING@CONNECT.UST.HK

**Daniel P. Palomar**

*Department of Electronic and Computer Engineering  
Department of Industrial Engineering and Decision Analytics  
The Hong Kong University of Science and Technology  
Clear Water Bay, Hong Kong*

PALOMAR@UST.HK

## Abstract

In the past two decades, the field of applied finance has tremendously benefited from graph theory. As a result, novel methods ranging from asset network estimation to hierarchical asset selection and portfolio allocation are now part of practitioners' toolboxes. In this paper, we investigate the fundamental problem of learning undirected graphical models under Laplacian structural constraints from the point of view of financial market times series data. In particular, we present natural justifications, supported by empirical evidence, for the usage of the Laplacian matrix as a model for the precision matrix of financial assets, while also establishing a direct link that reveals how Laplacian constraints are coupled to meaningful physical interpretations related to the market index factor and to conditional correlations between stocks. Those interpretations lead to a set of guidelines that practitioners should be aware of when estimating graphs in financial markets. In addition, we design numerical algorithms based on the alternating direction method of multipliers to learn undirected, weighted graphs that take into account stylized facts that are intrinsic to financial data such as heavy tails and modularity. We illustrate how to leverage the learned graphs into practical scenarios such as stock time series clustering and foreign exchange network estimation. The proposed graph learning algorithms outperform the state-of-the-art, benchmark methods in an extensive set of practical experiments, evidencing the advantages of adopting more principled assumptions into the learning framework. Furthermore, we obtain theoretical and empirical convergence results for the proposed algorithms. Along with the developed methodologies for graph learning in financial markets, we release an R package, called **fingraph** , accommodating the code and data to obtain all the experimental results.

**Keywords:** Graphs, Financial Markets, Quantitative Finance, Unsupervised Learning

## 1. Introduction

Graph learning from data has been a problem of critical importance for the statistical graph learning and graph signal processing fields (Friedman et al., 2008; Lake and Tenenbaum, 2010; Witten et al., 2011; Kalofolias, 2016; Egilmez et al., 2017; Pavez et al., 2018; Zhao et al., 2019), with direct impact on applied areas such as unsupervised learning, clustering (Hsieh et al., 2012; Sun et al., 2014; Tan et al., 2015; Nie et al., 2016; Hao et al., 2018; Kumar et al., 2020, 2019a), applied finance (Mantegna, 1999; de Prado, 2016; Marti et al., 2017a), network topology inference (Segarra et al., 2017; Mateos et al., 2019; Coutino et al., 2019; Shafipour and Mateos, 2020), community detection (Fortunato, 2010; Li et al., 2018; Chen et al., 2019), and graph neural nets (Wu et al., 2019; Pal et al., 2019).

The basic idea behind graph learning is to answer the following question: given a data matrix whose columns represent signals (observations) measured at the graph nodes, how can one design a graph representation that “best” fits such data matrix without possibly any (or with at most partial) knowledge of the underlying graph structure? By “graph representation” or “graph structure”, it is often understood the Laplacian, adjacency, or incidence matrices of the graph, or even a more general graph shift operator (Marques et al., 2016). In addition, the observed signals need not to live in regular, ordered spaces and can take arbitrary values, such as categorical and numerical, hence the probability distribution of the data can be highly unknown. Figure 1 illustrates such setting.

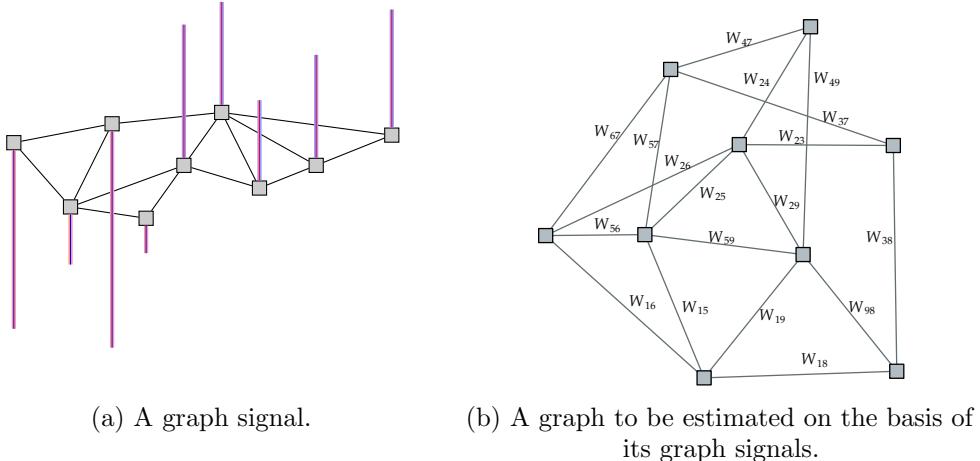


Figure 1: Illustration of a hypothetical signal observed in a graph (Figure 1a). The gray squares represent the graph nodes, the thin black lines denote the graph edges, indicating the relationships among nodes, whereas the vertical red bars denote the signal intensities measured at each node. Graph learning techniques seek to estimate the underlying graph structure (edge weights  $W_{ij}$  in Figure 1b) through the graph signal measurements.

Data derived from financial instruments such as equities and foreign exchanges, on the other hand, are defined on the well-known, ordered time domain (equally-spaced intraday<sup>1</sup>,

1. In high-frequency trading systems, in contrast, tick data may not necessarily be uniformly sampled. This scenario is not contemplated in this work.

daily, weekly, monthly, etc) and take on real values whose returns are often modeled by Gaussian processes. Then, the question is: How can graph learning be a useful tool for financial data analysis? In fact, [Mantegna \(1999\)](#) in his pioneer work showed that learning topological arrangements, such as graphs, in a stock market context, provides critical information that reveals economic factors that affect the price data evolution. The benefits that graph representations bring to applications on financial stocks are vastly discussed in the literature. A non-exhaustive, yet representative list of examples include: (i) identifying “business as usual” and “crash” periods via asset tree graphs ([Onnela et al., 2003b,a](#)), (ii) understanding portfolio dynamics via the topological properties of simulated minimum spanning trees ([Bonanno et al., 2003, 2004](#); [Mantegna and Stanley, 2004](#)), (iii) constructing networks of companies based on graphs ([Onnela et al., 2004](#)), (iv) understanding risks associated with a portfolio ([Malevergne and Sornette, 2006](#)), (v) leveraging properties of the learned graph into follow-up tasks such as hierarchical portfolio designs ([de Prado, 2016](#); [Raffinot, 2018a,b](#)), (vi) mining the relationship structure among investors ([Yang et al., 2020](#)), (vii) exploring graph properties such as degree centrality and eigenvector centrality for market crash detection and portfolio construction ([Millington and Niranjan, 2020](#)), and (viii) community detection in financial stock markets ([Ramakrishna et al., 2020](#); [de M. Cardoso and Palomar, 2020](#)).

Despite the plethora of applications, learning the structure of general graphical models is an NP-hard problem ([Anandkumar et al., 2012](#)) whose importance is critical towards visualizing, understanding, and leveraging the full potential contained in the data that live in such structures. Nonetheless, most existing techniques for learning graphs are often unable to impose a particular graph structure due to their inability to incorporate prior information in the learning process. More surprisingly, as it is shown in this work, state-of-the-art learning algorithms fall short when it comes to estimate graphs that posses certain properties such as  $k$ -components.

Moreover, graph learning frameworks are designed with the operational assumption that the observed graph signals are Gaussian distributed ([Friedman et al., 2008](#); [Lake and Tenenbaum, 2010](#); [Dong et al., 2016](#); [Kalofolias, 2016](#); [Egilmez et al., 2017](#); [Zhao et al., 2019](#); [Kumar et al., 2020](#); [Ying et al., 2020b](#)), inherently neglecting situations where there may exist outliers. As a consequence, those methods may not succeed in fully capturing a meaningful representation of the underlying graph especially in data from financial instruments, which are known to be heavy-tailed and skewed ([Gourioux and Monfort, 1997](#); [Cont, 2001](#); [Tsay, 2010](#); [Harvey, 2013](#); [Feng and Palomar, 2015](#); [Liu et al., 2019](#)).

While estimators for connected graphs have been proposed ([Dong et al., 2016](#); [Kalofolias, 2016](#); [Egilmez et al., 2017](#); [Zhao et al., 2019](#)), some of its properties, such as sparsity, are yet being investigated ([Ying et al., 2020b,a](#)), and only recently [Kumar et al. \(2020\)](#) have presented estimators for learning more general graphical structures such as  $k$ -component, bipartite, and  $k$ -component bipartite. However, one major shortcoming in ([Kumar et al., 2020](#)) is the lack of constraints on the degrees of the nodes. As we show in this work, the ability to control the degrees of the nodes is key to avoid trivial solutions while learning  $k$ -component graphs.

Recently, [Nie et al. \(2016\)](#); [Kumar et al. \(2020, 2019a\)](#) proposed optimization programs for learning the class of  $k$ -component graphs, as such class is an appealing model for clustering tasks due to the spectral properties of the Laplacian matrix. From a financial perspective,

clustering financial time-series, such as stock return data, has been an active research topic (Mantegna, 1999; Dose and Cincotti, 2005; Marti et al., 2016, 2017a,b). However, these works rely primarily on hierarchical clustering techniques and on the assumption that the underlying graph has a tree structure, which does bring advantages due to its hierarchical clustering properties, but also have been shown to be unstable (Carlsson and Mémoli, 2010; Lemieux et al., 2014; Marti et al., 2015) and not suitable when the data is not Gaussian distributed (Donnat et al., 2016). In this work, on the contrary, we tackle the problem of clustering stocks from a probabilistic perspective, similarly to the approach layed out by Kumar et al. (2019a, 2020), where the Laplacian matrix of a  $k$ -component graph is assumed to model the pairwise conditional correlations between stocks. A crucial advantage of this approach is that we can consider more realistic probabilistic assumptions such as heavy tails.

In practice, prior information about clusters of stocks is available via sector classification systems such as the Global Industry Classification Standard (GICS) (Standard & Poor's, 2006; Morgan Stanley Capital International and S&P Dow Jones, 2018) or the Industry Classification Benchmark (ICB) (Schreiner, 2019). However, more often than not, stocks have impacts on multiple industries, *e.g.*, the evident case of technology companies, such as Amazon, Apple, Google, and Facebook, whose influence on prices affect stocks not only in their own sector, but spans across multiple sectors. One reason for this phenomena is the myriad of services offered by those companies, resulting in challenges to precisely pin point which stock market sector they should belong to.

Motivated by practical applications in finance such as clustering of financial instruments and network estimation, we investigate the problem of learning graph matrices whose structure follow that of a Laplacian matrix of an undirected weighted graph.

The main contributions of our paper include:

1. As far as the authors are aware of, we for the first time provide interpretations for Laplacian constraints of graphs from the perspective of stock market data. Those interpretations naturally lead to meaningful and intuitive guidelines on the data pre-processing required for learning graphs from financial data.
2. We show that rank constraints alone, a practice often used by state-of-the-art methods, are not sufficient to learn non-trivial  $k$ -component graphs. We achieve learning of  $k$ -component graphs without isolated nodes by leveraging linear constraints on the node degrees of the graph.
3. We propose novel formulations to learn  $k$ -component graphs and heavy-tailed graphs, which are solved via carefully designed Alternating Direction Method of Multipliers (ADMM) algorithms. In addition, we establish theoretical convergence guarantees for the proposed algorithms along with experiments on their empirical convergence. The proposed algorithms can be easily extended to account for additional linear constraints on the graph weights.
4. We present extensive practical results that showcase the advantage of the operational assumptions used in the proposed algorithms when compared to state-of-the-art methods. Along with the methods proposed in this paper, we release an R package,

called **fingraph** , containing fast, unit-tested code that implements the proposed algorithms and it is publicly available at: <https://github.com/mirca/fingraph>.

## 1.1 Notation

The reals, nonnegative reals, and positive reals fields are denoted as  $\mathbb{R}$ ,  $\mathbb{R}_+$ , and  $\mathbb{R}_{++}$ , respectively. We use the abbreviation iff to denote “if and only if”. Scalars and real-valued random variables are denoted by lowercase roman letters like  $x$ . Matrices (vectors) are denoted by bold, italic, capital (lowercase) roman letters like  $\mathbf{X}$ ,  $\mathbf{x}$ . Vectors are assumed to be column vectors. The  $(i, j)$  element of a matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is denoted as  $X_{ij}$ . The  $i$ -th row (column) of  $\mathbf{X}$  is denoted as  $\mathbf{x}_{i,*} \in \mathbb{R}^{p \times 1}$  ( $\mathbf{x}_{*,i} \in \mathbb{R}^{n \times 1}$ ). The  $i$ -th element of a vector  $\mathbf{x}$  is denoted as  $x_i$ . The transpose of  $\mathbf{X}$  is denoted as  $\mathbf{X}^\top$ . The identity matrix of order  $p$  is denoted as  $\mathbf{I}_p$ . The Moore-Penrose inverse of a matrix  $\mathbf{X}$  is denoted as  $\mathbf{X}^\dagger$ . The trace of a square matrix, *i.e.*, the sum of elements on the principal diagonal, is denoted as  $\text{tr}(\mathbf{X})$ . The inner product between two matrices  $\mathbf{X}$ ,  $\mathbf{Y}$  is denoted as  $\langle \mathbf{X}, \mathbf{Y} \rangle \triangleq \text{tr}(\mathbf{X}^\top \mathbf{Y})$ . The element-wise sum of the absolute values and the Frobenius norm of a matrix  $\mathbf{X}$  are denoted as  $\|\mathbf{X}\|_1 = \sum_{ij} |X_{ij}|$  and  $\|\mathbf{X}\|_F = \sqrt{\text{tr}(\mathbf{X}^\top \mathbf{X})}$ , respectively. For  $\mathbf{x} \in \mathbb{R}^p$ ,  $\|\mathbf{x}\|_2$  stands for the usual Euclidean norm of  $\mathbf{x}$ . If  $\mathbf{A}$  is a symmetric matrix,  $\lambda_{\max}(\mathbf{A})$  denotes the maximum eigenvalue of  $\mathbf{A}$ . Let  $\mathbf{A}$ ,  $\mathbf{B}$  be two self-adjoint matrices. We write  $\mathbf{A} \succeq \mathbf{B}$  ( $\mathbf{A} \succ \mathbf{B}$ ) iff  $\mathbf{A} - \mathbf{B}$  is nonnegative (positive) definite. The symbols  $\mathbf{1}$  and  $\mathbf{0}$  denote the all ones and zeros vectors of appropriate dimension, respectively. The operator  $\text{diag} : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^p$  extracts the diagonal of a square matrix. The operator  $\text{Diag} : \mathbb{R}^p \rightarrow \mathbb{R}^{p \times p}$  creates a diagonal matrix with the elements of an input vector along its diagonal.  $(\mathbf{x})^+$  denotes the projection of  $\mathbf{x}$  onto the nonnegative orthant, *i.e.*, the elementwise maximum between  $\mathbf{0}$  and  $\mathbf{x}$ .

## 2. Background and Related Works

An undirected, weighted graph is usually denoted as a triple  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ , where  $\mathcal{V} = \{1, 2, \dots, p\}$  is the vertex (or node) set,  $\mathcal{E} \subseteq \{\{u, v\} : u, v \in \mathcal{V}\}$  is the edge set, that is, a subset of the set of all possible unordered pairs of  $p$  nodes such that  $\{u, v\} \in \mathcal{E}$  iff nodes  $u$  and  $v$  are connected.  $\mathbf{W} \in \mathbb{R}_+^{p \times p}$  is the symmetric weighted adjacency matrix that satisfies  $W_{ii} = 0$ ,  $W_{ij} > 0$  iff  $\{i, j\} \in \mathcal{E}$  and  $W_{ij} = 0$ , otherwise. We denote a graph as a 4-tuple  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W}, f_t)$ , where  $f_t : \mathcal{V} \rightarrow \{1, 2, \dots, t\}$  is a function that associates a single type (label) to each vertex of the graph, where  $t$  is the number of possible types. This extension is necessary for computing certain graph properties of practical interest such as graph modularity. We denote the number of elements in  $\mathcal{E}$  by  $|\mathcal{E}|$ . The combinatorial, unnormalized graph Laplacian matrix  $\mathbf{L}$  is defined, as usual, as  $\mathbf{L} \triangleq \mathbf{D} - \mathbf{W}$ , where  $\mathbf{D} \triangleq \text{Diag}(\mathbf{W}\mathbf{1})$  is the degree matrix.

An Improper Gaussian Markov Random Field (IGMRF) (Rue and Held, 2005; Slawski and Hein, 2015) of rank  $p - k$ ,  $k \geq 1$ , is denoted as a  $p$ -dimensional, real-valued, Gaussian random variable  $\mathbf{x}$  with mean vector  $\mathbb{E}[\mathbf{x}] \triangleq \boldsymbol{\mu}$  and rank-deficient precision matrix  $\mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]^\dagger \triangleq \boldsymbol{\Xi}$ . The probability density function of  $\mathbf{x}$  is then given as

$$p(\mathbf{x}) \propto \sqrt{\det^*(\boldsymbol{\Xi})} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Xi} (\mathbf{x} - \boldsymbol{\mu}) \right\}, \quad (1)$$

where  $\det^*(\boldsymbol{\Xi})$  is the pseudo (also known as generalized) determinant of  $\boldsymbol{\Xi}$ , *i.e.*, the product of its positive eigenvalues (Knill, 2014).

The data generating process is assumed to be a zero-mean, IGMRF  $\mathbf{x} \in \mathbb{R}^p$ , such that  $x_i$  is the random variable generating a signal measured at node  $i$ , whose rank-deficient precision matrix is modeled as a graph Laplacian matrix. This model is also known as Laplacian constrained Gaussian Markov Random Field (LGMRF) (Ying et al., 2020b). Assume we are given  $n$  observations of  $\mathbf{x}$ , *i.e.*,  $\mathbf{X} = [\mathbf{x}_{1,*}^\top, \mathbf{x}_{2,*}^\top, \dots, \mathbf{x}_{n,*}^\top]^\top$ ,  $\mathbf{X} \in \mathbb{R}^{n \times p}$ ,  $\mathbf{x}_{i,*} \in \mathbb{R}^{p \times 1}$ . The goal of graph learning algorithms is to learn a Laplacian matrix, or equivalently an adjacency matrix, given only the data matrix  $\mathbf{X}$ , *i.e.*, often without any knowledge of  $\mathcal{E}$  and  $f_t$ .

To that end, the classical penalized Maximum Likelihood Estimator (MLE) of the Laplacian-constrained precision matrix of  $\mathbf{x}$ , on the basis of the observed data  $\mathbf{X}$ , may be formulated as the following optimization program:

$$\begin{aligned} & \underset{\mathbf{L} \succeq 0}{\text{minimize}} \quad \text{tr}(\mathbf{LS}) - \log \det^*(\mathbf{L}) + h_{\boldsymbol{\alpha}}(\mathbf{L}), \\ & \text{subject to} \quad \mathbf{L}\mathbf{1} = \mathbf{0}, \quad L_{ij} = L_{ji} \leq 0, \end{aligned} \quad (2)$$

where  $\mathbf{S}$  is a similarity matrix, *e.g.*, the sample covariance (or correlation) matrix  $\mathbf{S} \propto \mathbf{X}^\top \mathbf{X}$ , and  $h_{\boldsymbol{\alpha}}(\mathbf{L})$  is a regularization function, with hyperparameter vector  $\boldsymbol{\alpha}$ , to promote certain properties on  $\mathbf{L}$ , such as sparsity or low-rankness.

Problem (2) is a fundamental problem in the graph signal processing field that has served as a cornerstone for many extensions, primarily those involving the inclusion of structure onto  $\mathbf{L}$  (Egilmez et al., 2017; Pavez et al., 2018; Kumar et al., 2019a,b, 2020). Even though Problem (2) is convex, provided we assume a convex choice for  $h_{\boldsymbol{\alpha}}(\cdot)$ , it is not adequate to be solved by Disciplined Convex Programming (DCP) languages, such as cvxpy (Diamond and Boyd, 2016), particularly due to scalability issues related to the computation of the term  $\log \det^*(\mathbf{L})$  (Egilmez et al., 2017; Zhao et al., 2019). Indeed, recently, considerable efforts have been directed towards the design of scalable, iterative algorithms based on Block Coordinate Descent (BCD) (Wright, 2015), Majorization-Minimization (MM) (Sun et al., 2017), and ADMM (Boyd et al., 2011) to solve Problem (2) in an efficient fashion, *e.g.*, (Egilmez et al., 2017), (Zhao et al., 2019), (Kumar et al., 2020), and (Ying et al., 2020b), just to name a few.

To circumvent some of those scalability issues related to the computation of the term  $\log \det^*(\mathbf{L})$ , Lake and Tenenbaum (2010) proposed the following relaxed version with an  $\ell_1$ -norm penalization:

$$\begin{aligned} & \underset{\tilde{\mathbf{L}} \succ 0, \mathbf{W}, \sigma > 0}{\text{maximize}} \quad -\text{tr}(\tilde{\mathbf{L}}\mathbf{S}) + \log \det(\tilde{\mathbf{L}}) - \alpha \|\mathbf{W}\|_1, \\ & \text{subject to} \quad \tilde{\mathbf{L}} = \text{Diag}(\mathbf{W}\mathbf{1}) - \mathbf{W} + \sigma \mathbf{I}_p, \\ & \quad \text{diag}(\mathbf{W}) = \mathbf{0}, \quad W_{ij} = W_{ji} \geq 0, \quad \forall i, j \in \{1, 2, \dots, p\}. \end{aligned} \quad (3)$$

In words, Problem (3) relaxes the original problem by forcing the precision matrix to be positive definite through the introduction of the term  $\sigma \mathbf{I}_p$ , which bounds the minimum eigenvalue of  $\tilde{\mathbf{L}}$  to be at least  $\sigma$ , and thus the generalized determinant can be replaced by the usual determinant. Although the technical issues related to the generalized determinant have been seemingly dealt with (albeit in an indirect way), in this formulation, there are twice as many variables to be estimated, which turns out to be prohibitive when designing

practical, scalable algorithms, and the applicability of `cvx`, like it was done in (Lake and Tenenbaum, 2010), is only possible in small scale ( $p \approx 50$ ) scenarios.

In order to solve this scalability issue, Hassan-Moghaddam et al. (2016) and Egilmez et al. (2017) proposed customized BCD algorithms (Shalev-Shwartz and Tewari, 2011; Saha and Tewari, 2013; Wright, 2015) to solve Problem (2) assuming an  $\ell_1$ -norm regularization, *i.e.*,  $h_\alpha(\mathbf{L}) \triangleq \alpha \sum_{i \neq j} |\mathbf{L}_{ij}| = -\alpha \sum_{i \neq j} \mathbf{L}_{ij}$ , in order to promote sparsity on the resulting estimated Laplacian matrix. However, as recently shown by Ying et al. (2020b,a), in contrast to common practices, the  $\ell_1$ -norm penalization surprisingly leads to denser graphs to the point that, for a large value of  $\alpha$ , the resulting graph will be fully connected with uniformly distributed graph weights.

On the other hand, due to such nuisances involved in dealing with the term  $\log \det^*(\mathbf{L})$ , several works departed from the LGMRF formulation altogether. Instead, they focused on the assumption that the underlying signals in a graph are smooth (Kalofolias, 2016; Dong et al., 2016; Chepuri et al., 2017). In its simplest form, learning a smooth graph from a data matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is tantamount to finding an adjacency matrix  $\mathbf{W}$  that minimizes the Dirichlet energy, *i.e.*,

$$\begin{aligned} & \underset{\mathbf{W}}{\text{minimize}} \quad \frac{1}{2} \sum_{i,j} W_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2, \\ & \text{subject to} \quad W_{ij} = W_{ji} \geq 0, \quad \text{diag}(\mathbf{W}) = \mathbf{0}. \end{aligned} \tag{4}$$

Problem (4) can also be equivalently expressed in terms of the Laplacian matrix:

$$\begin{aligned} & \underset{\mathbf{L} \succeq \mathbf{0}}{\text{minimize}} \quad \text{tr}(\mathbf{X} \mathbf{L} \mathbf{X}^\top), \\ & \text{subject to} \quad \mathbf{L} \mathbf{1} = \mathbf{0}, \quad L_{ij} = L_{ji} \leq 0. \end{aligned} \tag{5}$$

In order for Problems (4) and (5) to be well-defined, *i.e.*, to avoid the trivial solution  $\mathbf{W} = \mathbf{0}$  ( $\mathbf{L} = \mathbf{0}$ ), several constraints have been proposed in the literature. For instance, Dong et al. (2016) proposed one of the first estimators for graph Laplacian as the following nonconvex optimization program:

$$\begin{aligned} & \underset{\mathbf{L} \succeq \mathbf{0}, \mathbf{Y} \in \mathbb{R}^{n \times p}}{\text{minimize}} \quad \|\mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 + \alpha \text{tr}(\mathbf{Y} \mathbf{L} \mathbf{Y}^\top) + \eta \|\mathbf{L}\|_{\text{F}}^2, \\ & \text{subject to} \quad \mathbf{L} \mathbf{1} = \mathbf{0}, \quad L_{ij} = L_{ji} \leq 0, \quad \text{tr}(\mathbf{L}) = p, \end{aligned} \tag{6}$$

where the constraint  $\text{tr}(\mathbf{L}) = p$  is imposed to fix the sum of the degrees of the graph, and  $\alpha$  and  $\eta$  are positive, real-valued hyperparameters that control the amount of sparsity in the estimated graph. More precisely, for a given fixed  $\alpha$ , increasing (decreasing)  $\eta$  leads to sparser (denser) graphs. Dong et al. (2016) adopted an iterative alternating minimization scheme in order to find an optimal point of Problem (6), whereby at each iteration one of the variables is fixed while the solution is found for the other variable. However, the main shortcoming of formulation (6) is that it does not scale well for big data sets due to the update of the variable  $\mathbf{Y} \in \mathbb{R}^{n \times p}$ , which scales as a function of the number of observations  $n$ . In some graph learning problems in financial markets, the number of price recordings may be orders of magnitude larger than the number of nodes (financial assets), particularly in high frequency trading scenarios (Kirilenko et al., 2017).

Yet following the smooth signal assumption, [Kalofolias \(2016\)](#) proposed a convex formulation as follows:

$$\begin{aligned} & \underset{\mathbf{W}}{\text{minimize}} \quad \frac{1}{2} \text{tr}(\mathbf{WZ}) - \alpha \mathbf{1}^\top \log(\mathbf{W1}) + \frac{\eta}{2} \|\mathbf{W}\|_F^2, \\ & \text{subject to} \quad W_{ij} = W_{ji} \geq 0, \quad \text{diag}(\mathbf{W}) = \mathbf{0}, \end{aligned} \quad (7)$$

where  $Z_{ij} \triangleq \|\mathbf{x}_{*,i} - \mathbf{x}_{*,j}\|_2^2$ ,  $\alpha$  and  $\eta$  are positive, real-valued hyperparameters that control the amount of sparsity in the estimated graph, with the same interpretation as in Problem (6), and  $\log(\mathbf{W1})$ , assumed to be evaluated element-wise, is a regularization term added to avoid the degrees of the graph from becoming zero.

Problem (7) is convex and can be solved via primal-dual, ADMM-like algorithms ([Komodakis and Pesquet, 2015](#)). It can be seen that the objective function in Problem (7) is actually an approximation of that of Problem (2). From Hadamard's inequality ([Różański et al., 2017](#)), we have

$$\log \det^*(\mathbf{L}) \leq \log \prod_{i=1}^p L_{ii} = \sum_{i=1}^p \log L_{ii} = \mathbf{1}^\top \log(\mathbf{W1}). \quad (8)$$

Therefore, Problem (7) can be thought of as an approximation of the penalized maximum likelihood estimator with a Frobenius norm regularization in order to bound the graph weights.

The graph learning formulations previously discussed are only applicable to learn connected graphs. Learning graphs with a *prior* structure, *e.g.*,  $k$ -component graphs, poses a considerably higher challenge, as the dimension of the nullspace of the Laplacian matrix  $\mathbf{L}$  is equal to the number of components of the graph ([Chung, 1997](#)). Therefore, algorithms have to ensure that the algebraic multiplicity of the the 0 eigenvalue is equal to  $k$ . However, the latter condition is not sufficient to rule out the space of “trivial”  $k$ -component graphs, *i.e.*, graphs with isolated nodes. In addition to the rank (or nullity) constraint on  $\mathbf{L}$ , it is necessary to specify a constraint on the degrees of the graph.

Recent efforts have been made to introduce theoretical results from spectral graph theory ([Chung, 1997](#)) into practical optimization programs. For instance, a formulation to estimate  $k$ -component graphs based on the smooth signal approach was proposed in ([Nie et al., 2016](#)). More precisely, they proposed the Constrained Laplacian-rank (CLR) algorithm, which works in two-stages. On the first stage it estimates a connected graph using, *e.g.*, the solution to Problem (7), and then on the second stage it heuristically projects the graph onto the set of Laplacian matrices of dimension  $p$  with rank  $p - k$ , where  $k$  is the given number of graph components. This approach is summarized in the following two stages:

1. Obtain an initial affinity matrix  $\mathbf{A}^*$  as the optimal value of:

$$\begin{aligned} & \underset{\mathbf{A}}{\text{minimize}} \quad \frac{1}{2} \text{tr}(\mathbf{AZ}) + \frac{\eta}{2} \|\mathbf{A}\|_F^2, \\ & \text{subject to} \quad \text{diag}(\mathbf{A}) = \mathbf{0}, \quad \mathbf{A1} = \mathbf{1}, \quad A_{ij} \geq 0 \quad \forall i, j \end{aligned} \quad (9)$$

2. Find a projection of  $\mathbf{A}^*$  such that  $\mathbf{L}^* = \text{Diag}(\frac{\mathbf{B}^{*\top} + \mathbf{B}^*}{2}) - \frac{\mathbf{B}^{*\top} + \mathbf{B}^*}{2}$  has rank  $p - k$ :

$$\begin{aligned} & \underset{\mathbf{B}, \mathbf{L} \succeq 0}{\text{minimize}} \quad \|\mathbf{B} - \mathbf{A}^*\|_F^2, \\ & \text{subject to} \quad \mathbf{B1} = \mathbf{1}, \quad \text{rank}(\mathbf{L}) = p - k, \\ & \quad \mathbf{L} = \text{Diag}(\frac{\mathbf{B}^{\top} + \mathbf{B}}{2}) - \frac{\mathbf{B}^{\top} + \mathbf{B}}{2} \end{aligned} \quad (10)$$

where  $k$  is the desired number of graph components.

Spectral constraints on the Laplacian matrix are an intuitive way to recover  $k$ -component graphs as the multiplicity of its zero eigenvalue, *i.e.*, the nullity of  $\mathbf{L}$ , dictates the number of components of a graph. The first framework to impose structures on the estimated Laplacian matrix under the LGMRF model was proposed by Kumar *et al.* (Kumar et al., 2019a, 2020), through the use of spectral constraints, as follows:

$$\begin{aligned} & \underset{\mathbf{L}, \mathbf{U}, \boldsymbol{\lambda}}{\text{minimize}} \quad \text{tr}(\mathbf{L}\mathbf{S}) - \sum_{i=1}^{p-k} \log(\lambda_i) + \frac{\eta}{2} \left\| \mathbf{L} - \mathbf{U} \text{Diag}(\boldsymbol{\lambda}) \mathbf{U}^\top \right\|_{\text{F}}^2, \\ & \text{subject to} \quad \mathbf{L} \succeq \mathbf{0}, \quad \mathbf{L}\mathbf{1} = \mathbf{0}, \quad L_{ij} = L_{ji} \leq 0, \\ & \quad \mathbf{U}^\top \mathbf{U} = \mathbf{I}, \quad \mathbf{U} \in \mathbb{R}^{p \times (p-k)}, \\ & \quad \boldsymbol{\lambda} \in \mathbb{R}_+^{p-k}, \quad c_1 < \lambda_1 < \dots < \lambda_{p-k} < c_2. \end{aligned} \quad (11)$$

where the term  $\frac{\eta}{2} \left\| \mathbf{L} - \mathbf{U} \text{Diag}(\boldsymbol{\lambda}) \mathbf{U}^\top \right\|_{\text{F}}^2$ , often called spectral regularization, is added as a penalty term to indirectly promote  $\mathbf{L}$  to have the same rank as  $\mathbf{U} \text{Diag}(\boldsymbol{\lambda}) \mathbf{U}^\top$ , *i.e.*,  $p-k$ ,  $k$  is the number of components of the graph to be chosen *a priori*, and  $\eta > 0$  is a hyperparameter that controls the penalization on the spectral factorization of  $\mathbf{L}$ , and  $c_1$  and  $c_2$  are positive, real-valued constants employed to promote bounds on the eigenvalues of  $\mathbf{L}$ .

Note that Problem (11) learns a  $k$ -component graph without the need for a two-stage algorithm. However, a clear caveat of this formulation is that it does not control the degrees of the nodes in the graph, which may result in a trivial solution that contains isolated nodes, turning out not to be useful for clustering tasks especially when applied to noisy data sets or to data sets that are not significantly Gaussian distributed. In addition, choosing values for hyperparameters  $\eta$ ,  $c_1$ , and  $c_2$ , is often an intricate task.

### 3. Interpretations of Graph Laplacian Constraints for Financial Data

In this section, we present novel interpretations and motivations for the Laplacian constraints from the point of view of graphs learned from financial markets data. Those interpretations lead to paramount guidelines that users may benefit from when applying graph learning algorithms in financial problems. In addition, we provide sound justifications for the usage of the Laplacian matrix as a model for the inverse correlation matrix of financial assets.

Graphical representations of data are increasingly important tools in financial signal processing applied to uncover hidden relationships between variables (de Prado, 2016; Marti et al., 2017a). In financial markets, one is generally interested in learning quantifiable dependencies among assets and how to leverage them into practical scenarios such as portfolio design and crisis forecasting.

Arguably, one of the most successful methods to estimate sparse graphs is the Graphical Lasso (Friedman et al., 2008; Banerjee et al., 2008), modeled as the solution to the following convex optimization problem:

$$\underset{\boldsymbol{\Sigma}^{-1} \succ \mathbf{0}}{\text{minimize}} \quad \text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{S}) - \log \det(\boldsymbol{\Sigma}^{-1}) + \alpha \|\boldsymbol{\Sigma}^{-1}\|_1, \quad (12)$$

where  $\mathbf{S} \propto \mathbf{X}^\top \mathbf{X}$  is an empirical covariance (or correlation) matrix. The solution to Problem (12) can be efficiently computed via the well-known glasso algorithm (Friedman et al., 2008; Sustik and Calderhead, 2012).

While this model has been extremely successful in numerous fields, imposing a Laplacian structure onto  $\Sigma^{-1}$  brings significant benefits in financial data settings. To see that, we empirically evaluate the out-of-sample log-likelihood using three models:

1. Graphical Lasso as defined in (12).
2. Multivariate Totally Positive of Order 2 (MTP<sub>2</sub>), given as

$$\underset{\substack{\Sigma^{-1} \succ 0, \\ \Sigma_{ij}^{-1} \leq 0, \forall i \neq j}}{\text{minimize}} \quad \text{tr}(\Sigma^{-1} \mathbf{S}) - \log \det(\Sigma^{-1}) + \alpha \|\Sigma^{-1}\|_1. \quad (13)$$

3. Laplacian GMRF (LGMRF) as defined in (2), without regularization.<sup>2</sup>

We collect log-returns data from  $p = 50$  randomly chosen stocks from the S&P500 index during the period between Jan. 4th 2005 to Jul. 1st 2020 totalling  $n = 3900$  observations. We subdivide the observations into 26 sequential datasets each of which containing 150 observations. For the  $i$ -th dataset, we estimate the models (12) and (13) for different values of the hyperparameter  $\alpha$ , and compute their log-likelihood using the  $(i + 1)$ -th dataset. We then average the log-likelihood measurements over the datasets. In this fashion, we can infer how well these models generalize to unseen data.

Figure 2 shows the log-likelihood measurements in this experiment. We can readily notice that not only the LGMRF model has the higher explanatory power among the considered models, but it is also the simplest of them, as it does not contain any hyperparameter.

In addition, we plot one instance of the estimated networks from each of the models. Interestingly, Figure 3a reveals that Graphical Lasso estimates most conditional correlations as positive (blue edges), with only a few negative ones (red edges). In addition, both Graphical Lasso and MTP<sub>2</sub> (Figure 3b) do not clearly uncover strong connections between clearly correlated stocks. The LGMRF model (Figure 3c), on the other hand, displays vividly the interactions between evidently correlated nodes, *e.g.*, {CAH and MCK} and {SIVB and PBCT}, which are companies in the health care industry and bank holdings, respectively.

From the perspective of the LGMRF model, we would like to estimate a matrix  $\mathbf{L}$  that enjoys the following two key properties:

- (P1)  $\mathbf{L}\mathbf{1} = \mathbf{0}$ ,  
(P2)  $L_{ij} = L_{ji} \leq 0 \quad \forall i \neq j$ .

The first property states that the Laplacian matrix  $\mathbf{L}$  is singular and the eigenvector associated with its zero eigenvalue is given by  $a\mathbf{1}$ ,  $a \in \mathbb{R}$ , *i.e.*, the eigenvector is constant along all its components.

Based on the empirical and theoretical discussions about the spectrum of correlation matrices of stock time series (Plerou et al., 1999), we conduct an additional experiment to verify whether the sample inverse correlation matrix of stocks share the aforementioned properties: we query data from  $p = 414$  stocks belonging to the S&P500 index from January 4th 2005 to June 18th 2020, totalling  $n = 3869$  observations. We then divide this dataset

---

2. We do not regularize the LGMRF model with the  $\ell_1$ -norm as it leads to denser graphs (Ying et al., 2020a,b).

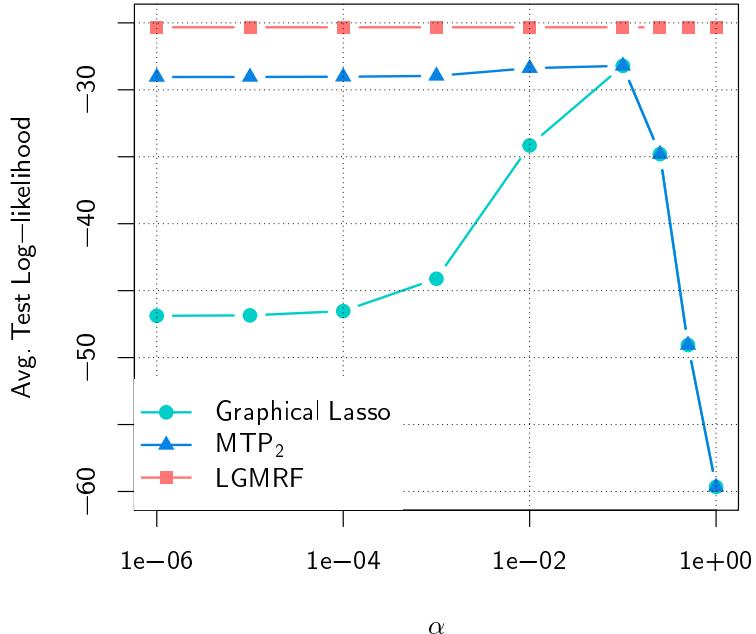


Figure 2: Average log-likelihood in out-of-sample data for different precision matrix estimation models as a function of the sparsity promoting hyperparameter  $\alpha$ .

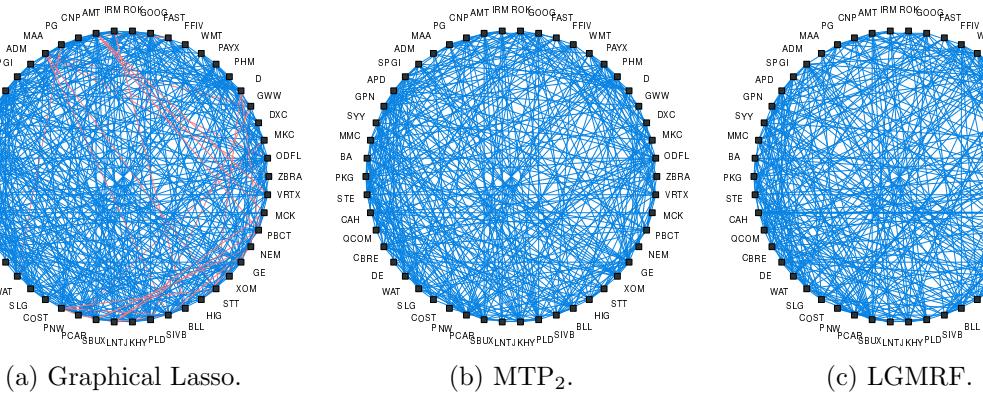


Figure 3: Estimated networks of stocks from (a) Graphical Lasso (b) MTP<sub>2</sub> and (c) LGMRF models at their highest average likelihood (Figure 2,  $\alpha = 10^{-1}$ ). The widths of the edges are proportional to the absolute value of the graph weights. Blue edges represent positive conditional correlations, while red edges represent negative ones.

into 19 sequential overlapping datasets each of which containing  $n = 2070$  observations, such that  $n/p = 5$ . For each dataset, we compute two attributes of the inverse sample correlation matrix: (i) its condition number, defined as the ratio between its maximum and minimum eigenvalues, and (ii) the variance of each eigenvector. We observe that the smallest condition number across all datasets is of the order of  $10^4$ , while the median of the condition numbers is of the order of  $10^5$ , indicating that in fact the inverse sample correlation matrices are nearly singular. In addition, the average variance of the eigenvector associated with the

zero eigenvalue is  $2 \cdot 10^{-4}$ , which is around an order of magnitude smaller than the average variances of any the other eigenvectors, indicating that it is, in fact, a constant eigenvector. Figure 4 illustrates this phenomenon for the aforementioned dataset, where the constant nature of the market eigenvector (eigenvector#1) is clearly observable when compared to the variability of the next two eigenvectors (eigenvector#2, eigenvector#3).

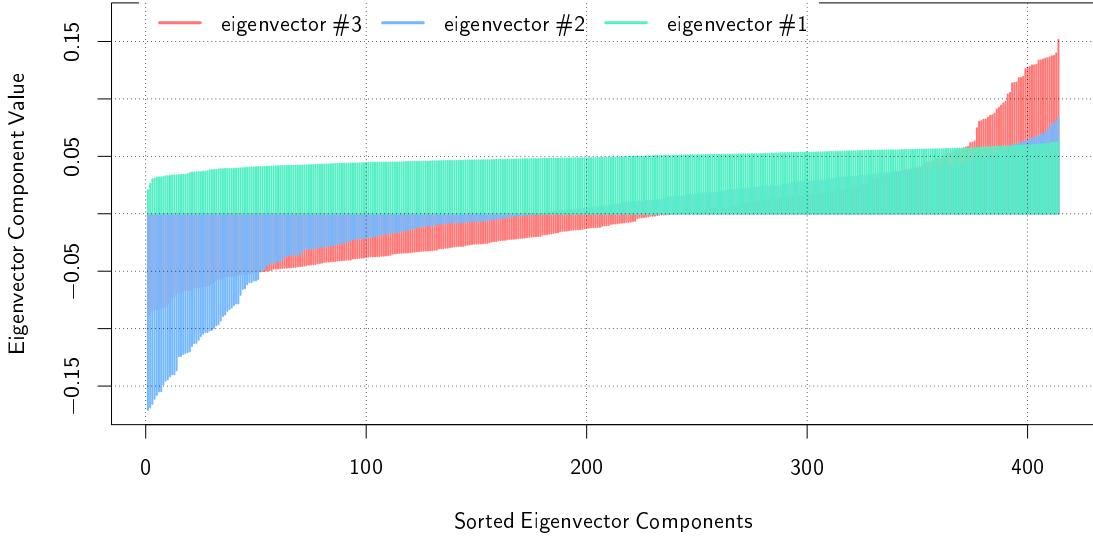


Figure 4: Eigenvectors of the sample correlation matrix of 414 S&P500 stocks corresponding to the largest three eigenvalues over the period between Jan. 2005 to Jun. 2020. eigenvector#1 represents the market factor. eigenvector#2 and #3 are displayed for reference on the expected variability.

In practice, (P1) implies that signals living in a graph  $\mathcal{G}$  have zero graph-mean, *i.e.*, the sum of the graph signals, at a given time, is zero. From a stock market perspective, the vector of log-returns of a set of stocks, at a given time  $i$ , is often assumed to follow a linear factor model (Sharpe, 1964; Fama and French, 2004), *i.e.*,  $\mathbf{x}_{*,i} = \boldsymbol{\beta}x_{\text{mkt},i} + \boldsymbol{\epsilon}_i$ , where  $\mathbf{x}_{*,i} \in \mathbb{R}^{p \times 1}$  contains the log-returns of  $p$  stocks,  $x_{\text{mkt},i} \in \mathbb{R}$  is the log-return of the market factor,  $\boldsymbol{\epsilon}_i$  is the vector of idiosyncratic log-returns that is often assumed to be a Gaussian process with zero mean vector and covariance matrix  $\boldsymbol{\Psi}$ , and  $\boldsymbol{\beta}$  is the vector of market factor loadings. Because  $\mathcal{G}$  has zero graph-mean, this implies that a graph designed to accommodate stock signals, assumed to follow a linear market factor model, will automatically remove the market component from the learning process of the conditional dependencies among stocks. This is a crucial feature because the market component would likely be a confounding factor in the estimation of the conditional correlations due to its strong influence on all the stocks log-returns.

In addition, (P2) together with (P1) implies that  $\mathbf{L}$  is positive semidefinite. The fact that the off-diagonal entries are symmetric and non-positive means that the Laplacian matrix only represents non-negative conditional dependencies<sup>3</sup>. This assumption is often met for

3. The correlation between any two pair of nodes conditioned on the rest of the graph is given as  $-\frac{L_{ij}}{\sqrt{L_{ii}L_{jj}}}$ .

stock data, as assets are typically positively dependent (Plerou et al., 2002; Kazakov and Kalyagin, 2016; Agrawal et al., 2020; Soloff et al., 2020; Wang et al., 2020).

These two properties along with efficient learning frameworks make the Laplacian-based graphical model a natural candidate for learning graphs of stock data. As a consequence of using the Laplacian model, we propose the following guidelines when estimating Laplacian matrices with stock market data:

- **Correlation vs Covariance:** Both the LGMRF and smooth signal approaches rely on the Dirichlet energy term  $\text{tr}(\mathbf{SL}) \propto \text{tr}(\mathbf{WZ})$ , which quantifies the smoothness of the graph signals over the graph weights, where  $\mathbf{S}$  is the sample covariance matrix. From the definition of  $\mathbf{Z}$  in (7), we observe that two perfectly correlated stocks but with large Euclidean distances would be translated as largely far apart nodes on the graph. Hence, we advocate the use of the sample correlation matrix  $\mathbf{S} = \text{Diag}(\mathbf{S})^{-1/2} \mathbf{S} \text{Diag}(\mathbf{S})^{-1/2}$  (or equivalently scaling the columns of  $\mathbf{X}$  such that they have unit variance) in case we would like two highly correlated stocks to have a strong graph connection regardless of their individual variances.
- **Removing the market trend:** A widely used and tested model for the returns of the stocks is the linear factor model, which explicitly includes the dependency on the market factor:  $\mathbf{x}_{*,i} = \beta x_{\text{mkt},i} + \boldsymbol{\epsilon}_i$ . Assuming that most of the stocks are heavily dominated by the market index  $x_{\text{mkt},i}$ , it may be convenient to remove that component if we seek to explore the structure of the residual cross-dependency among the stocks,  $\boldsymbol{\epsilon}_i$ . Thus, an alternative to using the full covariance matrix  $\boldsymbol{\Sigma}$  is to use the covariance matrix  $\boldsymbol{\Psi}$  of the idiosyncratic component. However, if one first normalizes each stock, whose variances are  $\mathbb{V}(\mathbf{x}_{*,i}) \approx \beta_i^2$ , we have  $\bar{\mathbf{x}}_{*,i} = \mathbf{1}\bar{x}_{\text{mkt},i} + \bar{\boldsymbol{\epsilon}}_i$ , then it turns out that the market factor is automatically removed in the normalized squared distance matrix  $\bar{\mathbf{Z}}$ :

$$Z_{ij} = \|\bar{\mathbf{x}}_{*,i} - \bar{\mathbf{x}}_{*,j}\|_2^2 = \|\bar{\boldsymbol{\epsilon}}_i - \bar{\boldsymbol{\epsilon}}_j\|_2^2. \quad (14)$$

- **Degree control:** Enforcing a rank smaller than  $p - 1$  on the Laplacian matrix will generate a  $k$ -component graph, which is one desired goal. However, one may get the undesired result of having isolated nodes. One possible strategy to avoid isolated nodes is via introducing constraints on the nodes degrees. The LGMRF formulation has the natural penalty term  $\log \det^*(\mathbf{L})$  in the objective, but that does not help in controlling the degrees of the nodes. Instead, some of the graph learning formulations from smooth signals include degree control via the constraint  $\mathbf{W1} = \mathbf{1}$ , which fixes the degrees of all the nodes to 1. The regularization term  $\mathbf{1}^\top \log(\mathbf{W1})$  also avoids the trivial solution of any degree equals 0. Hence, any graph learning formulation that enforces a  $k$ -component graph (or low-rank Laplacian matrix) should also control the degrees of the nodes to avoid a trivial solution with isolated nodes.

## 4. Proposed Algorithms

In this section, we design iterative algorithms for numerous graph learning formulations to account for  $k$ -component structures and the heavy-tail nature of financial stock market data.

The proposed algorithms are based on the ADMM (Boyd et al., 2011) and MM (Ortega and Rheinboldt, 2000; Sun et al., 2017) frameworks. We begin by briefly revisiting ADMM and MM.

#### 4.1 Alternating Direction Method of Multipliers (ADMM)

ADMM is a primal-dual framework designed to solve the following class of optimization problems:

$$\begin{aligned} & \underset{\mathbf{x}, \mathbf{z}}{\text{minimize}} \quad f(\mathbf{x}) + g(\mathbf{z}) \\ & \text{subject to} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}, \end{aligned} \quad (15)$$

where  $\mathbf{x} \in \mathbb{R}^n$  and  $\mathbf{z} \in \mathbb{R}^m$  are the optimization variables;  $\mathbf{A} \in \mathbb{R}^{p \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{p \times m}$ , and,  $\mathbf{c} \in \mathbb{R}^p$  are parameters; and  $f$  and  $g$  are convex, proper, closed, possibly non-differentiable functions.

The central object in the ADMM framework is the augmented Lagrangian function, which is given as

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\rho}{2} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|_2^2, \quad (16)$$

where  $\rho$  is a penalty parameter.

The basic workflow of the ADMM algorithm is summarized in Algorithm 1.

---

#### Algorithm 1: ADMM framework

---

**Data:**  $\mathbf{z}^0, \mathbf{y}^0, \mathbf{A}, \mathbf{B}, \mathbf{c}, \rho > 0$   
**Result:**  $\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*$

```

1  $l \leftarrow 0$ 
2 while not converged do
3    $\mathbf{x}^{l+1} \leftarrow \underset{\mathbf{x} \in \mathcal{X}}{\text{argmin}} L_\rho(\mathbf{x}, \mathbf{z}^l, \mathbf{y}^l)$ 
4    $\mathbf{z}^{l+1} \leftarrow \underset{\mathbf{z} \in \mathcal{Z}}{\text{argmin}} L_\rho(\mathbf{x}^{l+1}, \mathbf{z}, \mathbf{y}^l)$ 
5    $\mathbf{y}^{l+1} \leftarrow \mathbf{y}^l + \rho(\mathbf{A}\mathbf{x}^{l+1} + \mathbf{B}\mathbf{z}^{l+1} - \mathbf{c})$ 
6    $i \leftarrow l + 1$ 
7 end

```

---

The convergence of ADMM algorithms is attained provided that the following conditions are met:

1.  $\text{epi}(f) = \{(\mathbf{x}, t) \in \mathbb{R}^n \times \mathbb{R} : f(\mathbf{x}) \leq t\}$  and  $\text{epi}(g) = \{(\mathbf{z}, s) \in \mathbb{R}^m \times \mathbb{R} : g(\mathbf{z}) \leq s\}$  are both closed nonempty convex sets;
2. The unaugmented Lagrangian function  $L_0$  has a saddle point.

We refer readers to (Boyd et al., 2011) where elaborate convergence results are discussed.

#### 4.2 Majorization-Minimization (MM)

The MM framework seeks to solve the following general optimization problem:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad f(\mathbf{x}) \\ & \text{subject to} \quad \mathbf{x} \in \mathcal{X}, \end{aligned} \quad (17)$$

where here we consider  $f$  a smooth, possibly non-convex function.

The general idea behind MM is to find a sequence of feasible points  $\{\mathbf{x}^i\}_{i \in \mathbb{N}}$  by minimizing a sequence of carefully constructed global upper-bounds of  $f$ . The popular expectation-maximization (EM) algorithm is a special case of MM (Wu and Lange, 2010).

At point  $\mathbf{x}^i$ , we design a continuous global upper-bound function  $g(\cdot, \mathbf{x}^i) : \mathcal{X} \rightarrow \mathbb{R}$  such that

$$g(\mathbf{x}, \mathbf{x}^i) \geq f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \quad (18)$$

Then, in the minimization step we update  $\mathbf{x}$  as

$$\mathbf{x}^{i+1} \in \arg \min_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x}, \mathbf{x}^i). \quad (19)$$

The global upper-bound function  $g(\cdot, \mathbf{x}^i)$  must satisfy the following conditions in order to guarantee convergence:

1.  $g(\mathbf{x}, \mathbf{x}^i) \geq f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X},$
2.  $g(\mathbf{x}^i, \mathbf{x}^i) = f(\mathbf{x}^i),$
3.  $\nabla g(\mathbf{x}^i, \mathbf{x}^i) = \nabla f(\mathbf{x}^i),$
4.  $g(\mathbf{x}, \mathbf{x}^i)$  is continuous on both  $\mathbf{x}$  and  $\mathbf{x}^i$ .

A thorough discussion about MM, along with a significant number of its extensions, with practical examples, can be found in (Sun et al., 2017).

### 4.3 A Reformulation of the Graph Learning Problem

We formulate the graph learning problem from the LGMRF perspective as the following general optimization program:

$$\begin{aligned} & \underset{\mathbf{L} \succeq 0}{\text{minimize}} \quad \text{tr}(\mathbf{SL}) - \log \det^*(\mathbf{L}), \\ & \text{subject to} \quad \mathbf{L} \in \mathcal{C}_L, \quad \mathbf{L}\mathbf{1} = \mathbf{0}, \quad L_{ij} = L_{ji} \leq 0, \end{aligned} \quad (20)$$

where  $\mathcal{C}_L$  is a set describing additional constraints onto the structure of the estimated Laplacian matrix, *e.g.*,  $\mathcal{C}_L = \{\mathbf{L} : \text{diag}(\mathbf{L}) = d\mathbf{1}, d > 0\}$  specifies the set of  $d$ -regular graphs.

Now, to split the constraints in Problem (20), we introduce the following linear transformations: (a)  $\mathbf{L} = \mathcal{L}\mathbf{w}$ ,  $\mathbf{w} \in \mathbb{R}_+^{p(p-1)/2}$ , where  $\mathcal{L}$  is the Laplacian operator (*cf.* Definition (62)) and  $\mathbf{w}$  is the vector of edges weights; and (b)  $\Theta = \mathcal{L}\mathbf{w}$ . With this, we equivalently rewrite Problem (20) as

$$\begin{aligned} & \underset{\mathbf{w} \geq 0, \Theta \succeq 0}{\text{minimize}} \quad \text{tr}(\mathbf{SL}\mathbf{w}) - \log \det^*(\Theta), \\ & \text{subject to} \quad \Theta = \mathcal{L}\mathbf{w}, \quad \Theta \in \mathcal{C}_\Theta, \quad \mathbf{w} \in \mathcal{C}_w, \end{aligned} \quad (21)$$

where  $\mathcal{C}_\Theta$  and  $\mathcal{C}_w$  are sets describing additional constraints onto the structure of the estimated Laplacian matrix. For example, to estimate connected  $d$ -regular graphs we can use  $\mathcal{C}_\Theta = \{\Theta \in \mathbb{R}^{p \times p} : \text{rank}(\Theta) = p - 1\}$  together with  $\mathcal{C}_w = \{\mathbf{w} \in \mathbb{R}_+^{p(p-1)/2} : \mathbf{d}\mathbf{w} = d\mathbf{1}\}$ , where  $\mathbf{d}$  is the degree operator (*cf.* Definition (64)).

While Problem (21) can be convex for a limited family of graph structures, convex programming languages, such as `cvxpy`, have shown to perform poorly even for considerably small ( $p \approx 50$ ) graphs (Egilmez et al., 2017). Hence, we develop scalable algorithms based on the ADMM and MM frameworks.

#### 4.4 Connected Graphs

We first specialize Problem (21) to the class of connected graphs. The rationale for that is twofold: (1) while this problem has been well studied, we propose a significantly different algorithm than previous works (Egilmez et al., 2017; Zhao et al., 2019; Ying et al., 2020b) by splitting the optimization variables whereby additional constraints can be easily introduced and handled via ADMM; (2) in addition, the mathematical developments described for this simple class of graphs will serve as building blocks when we tackle more elaborate classes of graphs such as  $k$ -component or heavy-tailed.

For connected graphs, we rely on the fact that  $\det^*(\Theta) = \det(\Theta + \mathbf{J})$  (Egilmez et al., 2017), where  $\mathbf{J} = \frac{1}{p}\mathbf{1}\mathbf{1}^\top$ , to formulate the following convex optimization problem:

$$\begin{aligned} & \underset{\mathbf{w} \geq \mathbf{0}, \Theta \succeq \mathbf{0}}{\text{minimize}} \quad \text{tr}(\mathbf{S}\mathcal{L}\mathbf{w}) - \log \det(\Theta + \mathbf{J}), \\ & \text{subject to} \quad \Theta = \mathcal{L}\mathbf{w}, \quad \mathbf{d}\mathbf{w} = \mathbf{d}. \end{aligned} \quad (22)$$

The partial augmented Lagrangian function of Problem (22) can be written as

$$\begin{aligned} L_\rho(\Theta, \mathbf{w}, \mathbf{Y}, \mathbf{y}) &= \text{tr}(\mathbf{S}\mathcal{L}\mathbf{w}) - \log \det(\Theta + \mathbf{J}) + \langle \mathbf{y}, \mathbf{d}\mathbf{w} - \mathbf{d} \rangle + \frac{\rho}{2} \|\mathbf{d}\mathbf{w} - \mathbf{d}\|_2^2 \\ &+ \langle \mathbf{Y}, \Theta - \mathcal{L}\mathbf{w} \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L}\mathbf{w}\|_{\text{F}}^2, \end{aligned} \quad (23)$$

where  $\mathbf{Y}$  and  $\mathbf{y}$  are the dual variables associated with the constraints  $\Theta = \mathcal{L}\mathbf{w}$  and  $\mathbf{d}\mathbf{w} = \mathbf{d}$ , respectively. Note that we will deal with the constraints  $\mathbf{w} \geq \mathbf{0}$  and  $\Theta \succeq \mathbf{0}$  directly, hence there are no dual variables associated with them.

The subproblem for  $\Theta$  can be written as

$$\Theta^{l+1} = \arg \underset{\Theta \succeq \mathbf{0}}{\min} -\log \det(\Theta + \mathbf{J}) + \langle \Theta, \mathbf{Y}^l \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L}\mathbf{w}^l\|_{\text{F}}^2. \quad (24)$$

Now, making the simple affine transformation  $\Omega^{l+1} = \Theta^{l+1} + \mathbf{J}$ , we have

$$\Omega^{l+1} = \arg \underset{\Omega \succ \mathbf{0}}{\min} -\log \det(\Omega) + \langle \Omega, \mathbf{Y}^l \rangle + \frac{\rho}{2} \|\Omega - \mathcal{L}\mathbf{w}^l - \mathbf{J}\|_{\text{F}}^2, \quad (25)$$

which can be expressed as a proximal operator (Parikh and Boyd, 2014), *cf.* Definition (69),

$$\Omega^{l+1} = \text{prox}_{\rho^{-1}(-\log \det(\cdot) + \langle \mathbf{Y}^l, \cdot \rangle)}(\mathcal{L}\mathbf{w}^l + \mathbf{J}), \quad (26)$$

whose closed-form solution is given by Lemma 1.

**Lemma 1** *The global minimizer of problem (26) is (Witten and Tibshirani, 2009; Danaher et al., 2014)*

$$\Omega^{l+1} = \frac{1}{2\rho} \mathbf{U} \left( \mathbf{\Gamma} + \sqrt{\mathbf{\Gamma}^2 + 4\rho\mathbf{I}} \right) \mathbf{U}^\top, \quad (27)$$

where  $\mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top$  is the eigenvalue decomposition of  $\rho(\mathcal{L}\mathbf{w}^l + \mathbf{J}) - \mathbf{Y}^l$ .

Hence the closed-form solution for (24) is

$$\Theta^{l+1} = \Omega^{l+1} - J. \quad (28)$$

Now, using the linear properties of adjoint operators, we have that  $\text{tr}(S\mathcal{L}\mathbf{w}) = \langle \mathbf{w}, \mathcal{L}^*S \rangle$  and  $\|\mathcal{L}\mathbf{w}\|_F^2 = \text{tr}(\mathcal{L}\mathbf{w}\mathcal{L}\mathbf{w}) = \mathbf{w}^\top \mathcal{L}^* \mathcal{L} \mathbf{w}$ . Then, the subproblem for  $\mathbf{w}$  can be written as

$$\mathbf{w}^{l+1} = \arg \min_{\mathbf{w} \geq 0} \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} + \left\langle \mathbf{w}, \mathcal{L}^* (S - \mathbf{Y}^l - \rho \Theta^{l+1}) + \mathbf{d}^* (\mathbf{y}^l - \rho \mathbf{d}) \right\rangle, \quad (29)$$

which is a nonnegative, convex quadratic program.

**Lemma 2** *Problem (29) is strictly convex.*

**Proof** It suffices to show that the matrix  $\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}$  is positive definite. For any  $\mathbf{x} \in \mathbb{R}^{p(p-1)/2}$ ,  $\mathbf{x} \neq 0$ , we have that  $\|\mathbf{d}\mathbf{x}\|_F^2 = \langle \mathbf{d}\mathbf{x}, \mathbf{d}\mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{d}^* \mathbf{d} \mathbf{x} \rangle \geq 0$ . To see that  $\mathcal{L}^* \mathcal{L}$  is positive definite, we refer the readers to (Ying et al., 2020a, Lemma 5.3). ■

While the solution to Problem (29) might seem straightforward to obtain via quadratic programming solvers, it actually poses an insurmountable scalability issue: the dimension of the matrices  $\mathbf{d}^* \mathbf{d}$  and  $\mathcal{L}^* \mathcal{L}$  is  $p(p-1)/2 \times p(p-1)/2$ , implying that the worst-case complexity of a convex QP solver for this problem is  $O(p^6)$  (Ye and Tse, 1989), which is impractical. In addition, no closed-form solution is available.

Given these difficulties, we resort to the MM method, whereby we construct an upper-bound of the objective function of (29) at point  $\mathbf{w}^i = \mathbf{w}^l \in \mathbb{R}_+^{p(p-1)/2}$  as

$$g(\mathbf{w}, \mathbf{w}^i) = g(\mathbf{w}^i, \mathbf{w}^i) + \langle \mathbf{w} - \mathbf{w}^i, \nabla_{\mathbf{w}} f(\mathbf{w}^i) \rangle + \frac{\mu}{2} \|\mathbf{w} - \mathbf{w}^i\|_2^2, \quad (30)$$

where  $f(\cdot)$  is the objective function in the minimization in (29),  $\mu = \rho \lambda_{\max}(\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L})$ , and the maximum eigenvalue of  $\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}$  is given by Lemma 3.

**Lemma 3** *The maximum eigenvalue of the matrix  $\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}$  is given as*

$$\lambda_{\max}(\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) = 2(2p - 1). \quad (31)$$

**Proof** The proof is deferred to Appendix C.1. ■

Finally, we have that  $\nabla_{\mathbf{w}} f(\mathbf{w}^i) = \mathbf{a}^i + \mathbf{b}^i$ , where

$$\mathbf{a}^i = \mathcal{L}^* (S - \mathbf{Y}^l - \rho (\Theta^{l+1} - \mathcal{L}\mathbf{w}^i)), \quad (32)$$

$$\mathbf{b}^i = \mathbf{d}^* (\mathbf{y}^l - \rho (\mathbf{d} - \mathbf{d}\mathbf{w}^i)). \quad (33)$$

Thus, we have the following approximate strictly convex subproblem for  $\mathbf{w}$ ,

$$\mathbf{w}^{i+1} = \arg \min_{\mathbf{w} \geq 0} \rho(2p - 1) \|\mathbf{w} - \mathbf{w}^i\|_2^2 + \langle \mathbf{w}, \mathbf{a}^i + \mathbf{b}^i \rangle, \quad (34)$$

whose solution can be readily obtained via its KKT optimality conditions and its given as

$$\mathbf{w}^{i+1} = \left( \mathbf{w}^i - \frac{\mathbf{a}^i + \mathbf{b}^i}{2\rho(2p-1)} \right)^+, \quad (35)$$

which is a projected gradient descent step with learning rate  $2\rho(2p-1)$ . Thus, we iterate (35) in order to obtain the unique optimal point,  $\mathbf{w}^{l+1}$ , of Problem (29). In practice, we observe that a few ( $\approx 5$ ) iterations are sufficient for convergence.

The dual variables  $\mathbf{Y}$  and  $\mathbf{y}$  are updated as

$$\mathbf{Y}^{l+1} = \mathbf{Y}^l + \rho \left( \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1} \right) \quad (36)$$

and

$$\mathbf{y}^{l+1} = \mathbf{y}^l + \rho \left( \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \right). \quad (37)$$

A practical implementation for the proposed ADMM estimation of connected graphs is summarized in Algorithm 2, whose convergence is stated in Theorem 4.

---

**Algorithm 2:** Connected graph learning

---

**Data:** Similarity matrix  $\mathbf{S}$ , initial estimate of the graph weights  $\mathbf{w}^0$ , desired degree vector  $\mathbf{d}$ , penalty parameter  $\rho > 0$ , tolerance  $\epsilon > 0$   
**Result:** Laplacian estimation:  $\mathcal{L}\mathbf{w}^*$

```

1 initialize  $\mathbf{Y} = \mathbf{0}$ ,  $\mathbf{y} = \mathbf{0}$ 
2  $l \leftarrow 0$ 
3 while  $\max(|\mathbf{r}^l|) > \epsilon$  or  $\max(|\mathbf{s}^l|) > \epsilon$  do
4    $\triangleright$  update  $\Theta^{l+1}$  via (28)
5    $\triangleright$  iterate (35) until convergence so as to obtain  $\mathbf{w}^{l+1}$ 
6    $\triangleright$  update  $\mathbf{Y}^{l+1}$  as in (36)
7    $\triangleright$  update  $\mathbf{y}^{l+1}$  as in (37)
8    $\triangleright$  compute residual  $\mathbf{r}^{l+1} = \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}$ 
9    $\triangleright$  compute residual  $\mathbf{s}^{l+1} = \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}$ 
10   $l \leftarrow l + 1$ 
11 end

```

---

**Theorem 4** *The sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 2 converges to the optimal primal-dual solution of Problem (22).*

**Proof** The proof is deferred to Appendix C.2. ■

#### 4.5 $k$ -component Graphs

As discussed in Section 3, in addition to a rank constraint, some form of control of the node degrees is necessary to learn meaningful  $k$ -component graphs. Here we choose

$\mathcal{C}_L = \{L : \text{diag}(L) = d, d \in \mathbb{R}_{++}^p, \text{rank}(L) = p - k\}$ , which is translated to the framework of Problem (21) as

$$\mathcal{C}_\Theta = \{\Theta \succeq 0 : \text{rank}(\Theta) = p - k\}, \quad (38)$$

$$\mathcal{C}_w = \{w : \mathbf{d}w = d, \text{rank}(\mathcal{L}w) = p - k, d \in \mathbb{R}_{++}^p\}. \quad (39)$$

*Remark:* Although the rank constraints on both variables  $\Theta$  and  $w$  may seem redundant, we have observed that it greatly improves the empirical convergence of the algorithm. In addition, the rank constraint on  $\Theta$  does not incur any additional computational cost, as will be shown in the numerical algorithmic derivations below.

Thus, Problem (21) can be specialized for the task of learning a  $k$ -component graph as the following non-convex optimization program:

$$\begin{aligned} & \underset{w \geq 0, \Theta \succeq 0}{\text{minimize}} \quad \text{tr}(S\mathcal{L}w) - \log \det^*(\Theta), \\ & \text{subject to} \quad \Theta = \mathcal{L}w, \text{rank}(\Theta) = p - k, \mathbf{d}w = d, \text{rank}(\mathcal{L}w) = p - k. \end{aligned} \quad (40)$$

However, unlike the rank constraint in the subproblem associated with  $\Theta$ , the constraint  $\text{rank}(\mathcal{L}w) = p - k$  cannot be directly dealt with. An alternative is to move this constraint to the objective function by approximating it by noting that it is equivalent to having the sum of the  $k$  smallest eigenvalues of  $\mathcal{L}w$  equals zero, *i.e.*,  $\sum_{i=1}^k \lambda_i(\mathcal{L}w) = 0$  (Nie et al., 2016), assuming the sequence of eigenvalues  $\{\lambda_i(\mathcal{L}w)\}_{i=1}^p$  in increasing order. By Fan's theorem (Fan, 1949), we have

$$\sum_{i=1}^k \lambda_i(\mathcal{L}w) = \underset{\mathbf{V} \in \mathbb{R}^{p \times k}, \mathbf{V}^\top \mathbf{V} = \mathbf{I}}{\text{minimize}} \text{tr}(\mathbf{V}^\top \mathcal{L}w \mathbf{V}). \quad (41)$$

Thus, moving (41) into the objective function of Problem (40), we have the following relaxed problem:

$$\begin{aligned} & \underset{w \geq 0, \Theta, \mathbf{V}}{\text{minimize}} \quad \text{tr}(\mathcal{L}w(S + \eta \mathbf{V} \mathbf{V}^\top)) - \log \det^*(\Theta), \\ & \text{subject to} \quad \Theta = \mathcal{L}w, \text{rank}(\Theta) = p - k, \mathbf{d}w = d, \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \mathbf{V} \in \mathbb{R}^{p \times k}, \end{aligned} \quad (42)$$

where  $\eta > 0$  is a hyperparameter that controls how much importance is given to the term  $\text{tr}(\mathbf{V}^\top \mathcal{L}w \mathbf{V})$ , which indirectly promotes  $\text{rank}(\mathcal{L}w) = p - k$ . Therefore, via (41), we are able to incorporate the somewhat intractable constraint  $\text{rank}(\mathcal{L}w) = p - k$  as a simple term in the optimization program.

The partial augmented Lagrangian function of Problem (42) can be written as

$$\begin{aligned} L_\rho(\Theta, w, \mathbf{V}, \mathbf{Y}, \mathbf{y}) &= \text{tr}(\mathcal{L}w(S + \eta \mathbf{V} \mathbf{V}^\top)) - \log \det^*(\Theta) + \langle \mathbf{y}, \mathbf{d}w - d \rangle + \frac{\rho}{2} \|\mathbf{d}w - d\|_2^2 \\ &+ \langle \mathbf{Y}, \Theta - \mathcal{L}w \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L}w\|_F^2. \end{aligned} \quad (43)$$

The subproblem for  $\Theta$  can be written as

$$\Theta^{l+1} = \underset{\substack{\text{rank}(\Theta) = p - k \\ \Theta \succeq 0}}{\arg \min} - \log \det^*(\Theta) + \langle \Theta, \mathbf{Y}^l \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L}w^l\|_F^2, \quad (44)$$

which is tantamount to that of (25). Its solution is also given as

$$\Theta^* = \frac{1}{2\rho} \mathbf{U} \left( \mathbf{\Gamma} + \sqrt{\mathbf{\Gamma}^2 + 4\rho\mathbf{I}} \right) \mathbf{U}^\top, \quad (45)$$

except that now  $\mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top$  is the eigenvalue decomposition of  $\rho\mathcal{L}\mathbf{w}^l - \mathbf{Y}^l$ , with  $\mathbf{\Gamma}$  having the largest  $p - k$  eigenvalues along its diagonal and  $\mathbf{U} \in \mathbb{R}^{p \times (p-k)}$  contains the corresponding eigenvectors.

The update for  $\mathbf{w}$  is carried out similarly to that of (35), *i.e.*,

$$\mathbf{w}^{i+1} = \left( \mathbf{w}^i - \frac{\mathbf{a}^i + \mathbf{b}^i}{2\rho(2p-1)} \right)^+, \quad (46)$$

except that the coefficient  $\mathbf{a}^i$  is given as

$$\mathbf{a}^i = \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l \mathbf{V}^{l\top} - \mathbf{Y}^l - \rho \left( \Theta^{l+1} - \mathcal{L}\mathbf{w}^i \right) \right). \quad (47)$$

We have the following subproblem for  $\mathbf{V}$ :

$$\begin{aligned} & \underset{\mathbf{V} \in \mathbb{R}^{p \times k}}{\text{minimize}} \quad \text{tr}(\mathbf{V}^\top \mathcal{L}\mathbf{w}^{l+1} \mathbf{V}), \\ & \text{subject to} \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \end{aligned} \quad (48)$$

whose closed-form solution is given by the  $k$  eigenvectors associated with the  $k$  smallest eigenvalues of  $\mathcal{L}\mathbf{w}^{l+1}$  (Horn and Johnson, 1985; Absil et al., 2007).

The updates for the dual variables  $\mathbf{Y}$  and  $\mathbf{y}$  are exactly the same as in (36) and (37), respectively.

A practical implementation for the proposed ADMM estimation of  $k$ -component graphs is summarized in Algorithm 3, named kGL. Its complexity is bounded by the complexity of the eigenvalue decomposition in line 4. Its convergence is stated in Theorem 5.

**Theorem 5** *Algorithm 3 subsequently converges for any sufficiently large  $\rho$ , that is, the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 3 has at least one limit point, and each limit point is a stationary point of (43).*

**Proof** The proof is deferred to Appendix C.3. ■

#### 4.6 Connected heavy-tailed graphs

Following the LGMRF framework, Ying et al. (2020a,b) recently proposed non-convex regularizations so as to obtain sparse representations of the resulting estimated graphs. Enforcing sparsity is one possible way to remove spurious conditional correlations between nodes in the presence of data with outliers. However, we advocate that assuming a principled, heavy-tailed statistical distribution has more benefits for the financial data setting, rather than simply imposing arbitrary, non-convex regularizations, because they are often cumbersome to deal with from a theoretical perspective and, in practice, they bring the additional task of tuning hyperparameters, which is often repetitive.

---

**Algorithm 3:**  $k$ -component graph learning (kGL)
 

---

**Data:** Similarity matrix  $\mathbf{S}$ , initial estimate of the graph weights  $\mathbf{w}^0$ , desired number of graph components  $k$ , desired degree vector  $\mathbf{d}$ , penalty parameter  $\rho > 0$ , tolerance  $\epsilon > 0$

**Result:** Laplacian estimation:  $\mathcal{L}\mathbf{w}^*$

```

1 initialize  $\mathbf{Y} = \mathbf{0}$ ,  $\mathbf{y} = \mathbf{0}$ 
2  $l \leftarrow 0$ 
3 while  $\max(|\mathbf{r}^l|) > \epsilon$  or  $\max(|\mathbf{s}^l|) > \epsilon$  do
4    $\triangleright$  update  $\Theta^{l+1}$  via (45)
5    $\triangleright$  iterate (46) until convergence with  $\mathbf{a}^i$  given as in (47) so as to obtain  $\mathbf{w}^{l+1}$ 
6    $\triangleright$  update  $\mathbf{V}^{l+1}$  as in (48)
7    $\triangleright$  update  $\mathbf{Y}^{l+1}$  as in (36)
8    $\triangleright$  update  $\mathbf{y}^{l+1}$  as in (37)
9    $\triangleright$  compute residual  $\mathbf{r}^{l+1} = \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}$ 
10   $\triangleright$  compute residual  $\mathbf{s}^{l+1} = \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}$ 
11   $l \leftarrow l + 1$ 
12 end

```

---

In order to address the inherently heavy-tailed nature of financial stock data (Resnick, 2007), we consider the Student-t distribution under the Improper Markov Random Field assumption (Rue and Held, 2005) with Laplacian structural constraints, that is, we assume the data generating process to be modeled a multivariate zero-mean Student-t distribution, whose probability density function can be written as

$$p(\mathbf{x}) \propto \sqrt{\det^*(\Theta)} \left( 1 + \frac{\mathbf{x}^\top \Theta \mathbf{x}}{\nu} \right)^{-\frac{\nu+p}{2}}, \quad \nu > 2, \quad (49)$$

where  $\Theta$  is a positive-semidefinite inverse scatter matrix modeled as a combinatorial graph Laplacian matrix.

This results in a robustified version of the penalized MLE for connected graph learning, *i.e.*,

$$\begin{aligned} & \underset{\mathbf{w} \geq \mathbf{0}, \Theta > \mathbf{0}}{\text{minimize}} \quad \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L}\mathbf{w} \mathbf{x}_{i,*}}{\nu} \right) - \log \det(\Theta + \mathbf{J}), \\ & \text{subject to} \quad \Theta = \mathcal{L}\mathbf{w}, \quad \mathbf{d}\mathbf{w} = \mathbf{d}. \end{aligned} \quad (50)$$

Problem (50) is non-convex due to the terms involving the  $\log(\cdot)$  function and hence it is difficult to be dealt with directly. To tackle this issue, we leverage the MM framework whereby the concave terms in (50) are linearized, which essentially results in a weighted Gaussian likelihood (Sun et al., 2016, 2017; Wald et al., 2019).

We start by following the exposition in the preceding sections, then the partial augmented Lagrangian function of Problem (50) is given as

$$\begin{aligned} L_\rho(\Theta, \mathbf{w}, \mathbf{Y}, \mathbf{y}) &= \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right) - \log \det(\Theta + \mathbf{J}) + \langle \mathbf{y}, \mathbf{d} \mathbf{w} - \mathbf{d} \rangle \\ &\quad + \frac{\rho}{2} \|\mathbf{d} \mathbf{w} - \mathbf{d}\|_2^2 + \langle \mathbf{Y}, \Theta - \mathcal{L} \mathbf{w} \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L} \mathbf{w}\|_{\text{F}}^2. \end{aligned} \quad (51)$$

The subproblem for  $\Theta$  is identical to that of (24).

The subproblem for  $\mathbf{w}$  can be written as

$$\begin{aligned} \underset{\mathbf{w} \geq 0}{\text{minimize}} \quad & \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} - \left\langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{Y}^l + \rho \Theta^{l+1} \right) - \mathbf{d}^* \left( \mathbf{y}^l - \rho \mathbf{d} \right) \right\rangle \\ & + \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right), \end{aligned} \quad (52)$$

which is similar to that of subproblem (29), except it contains the additional concave term  $\frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right)$  in place of the linear term  $\langle \mathbf{S}, \mathcal{L} \mathbf{w} \rangle$ .

Similarly to subproblem (29), we employ the MM framework to formulate an efficient iterative algorithm to obtain a stationary point of Problem (52). We proceed by constructing a global upper bound of Problem (52). Using the fact that the logarithm is globally upper-bounded by its first-order Taylor expansion, we have

$$\log \left( 1 + \frac{t}{b} \right) \leq \log \left( 1 + \frac{a}{b} \right) + \frac{t - a}{a + b}, \quad \forall a \geq 0, t \geq 0, b > 2, \quad (53)$$

which results in the following upper bound:

$$\log \left( 1 + \frac{\langle \mathbf{w}, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle}{\nu} \right) \leq \frac{\langle \mathbf{w}, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle}{\langle \mathbf{w}^j, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle + \nu} + c_1 \quad (54)$$

where  $c_1 = \log \left( 1 + \frac{\langle \mathbf{w}^j, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle}{\nu} \right) - \frac{\langle \mathbf{w}^j, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle}{\langle \mathbf{w}^j, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle + \nu}$  is a constant.

By upper-bounding the objective function of Problem (52), at point  $\mathbf{w}^j = \mathbf{w}^l$ , with (54), the vector of graph weights  $\mathbf{w}$  can then be updated by solving the following nonnegative, quadratic-constrained, strictly convex problem:

$$\begin{aligned} \mathbf{w}^{j+1} &= \underset{\mathbf{w} \geq 0}{\arg \min} \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} - \left\langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{Y}^l + \rho \Theta^{l+1} \right) - \mathbf{d}^* \left( \mathbf{y}^l - \rho \mathbf{d} \right) \right\rangle \\ &\quad + \frac{p + \nu}{n} \sum_{i=1}^n \frac{\langle \mathbf{w}, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle}{\langle \mathbf{w}^j, \mathcal{L}^* \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top \rangle + \nu} \\ &= \underset{\mathbf{w} \geq 0}{\arg \min} \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} + \left\langle \mathbf{w}, \mathcal{L}^* \left( \tilde{\mathbf{S}}^j - \mathbf{Y}^l - \rho \Theta^{l+1} \right) + \mathbf{d}^* \left( \mathbf{y}^l - \rho \mathbf{d} \right) \right\rangle, \end{aligned} \quad (55)$$

where  $\tilde{\mathbf{S}}^j \triangleq \frac{1}{n} \sum_{i=1}^n \frac{(p + \nu)}{\langle \mathbf{w}^j, \mathcal{L}^*(\mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top) \rangle + \nu} \mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top$  is a weighted sample covariance matrix.

The objective function of Problem (55) can be upper-bounded once again following the same steps as the ones taken for Problem (30), which results in a projected gradient descent step as in (35) with

$$\mathbf{a}^j \triangleq \mathcal{L}^*(\tilde{\mathbf{S}}^j - \mathbf{Y}^l - \rho(\Theta^{l+1} - \mathcal{L}\mathbf{w}^j)). \quad (56)$$

The dual variables  $\mathbf{Y}$  and  $\mathbf{y}$  are updated exactly as in (36) and (37), respectively.

Algorithm 4, named tGL, summarizes the implementation to solve Problem (50). The complexity of Algorithm 4 is bounded by the complexity of the eigenvalue decomposition in line 4 and its convergence is stated by Theorem 6.

---

**Algorithm 4:** Connected Student- $t$  graph learning (tGL)

---

**Data:** Data matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$ , initial estimate of the graph weights  $\mathbf{w}^0$ , desired degree vector  $\mathbf{d}$ , penalty parameter  $\rho > 0$ , degrees of freedom  $\nu$ , tolerance  $\epsilon > 0$

**Result:** Laplacian estimation:  $\mathcal{L}\mathbf{w}^*$

```

1 initialize  $\mathbf{Y} = \mathbf{0}$ ,  $\mathbf{y} = \mathbf{0}$ 
2  $l \leftarrow 0$ 
3 while  $\max(|\mathbf{r}^l|) > \epsilon$  or  $\max(|\mathbf{s}^l|) > \epsilon$  do
4    $\triangleright$  update  $\Theta^{l+1}$  via (28)
5    $\triangleright$  iterate (35) with  $\mathbf{a}^j$  given as in (56) so as to obtain  $\mathbf{w}^{l+1}$ 
6    $\triangleright$  update  $\mathbf{Y}^{l+1}$  as in (36)
7    $\triangleright$  update  $\mathbf{y}^{l+1}$  as in (37)
8    $\triangleright$  compute residual  $\mathbf{r}^{l+1} = \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}$ 
9    $\triangleright$  compute residual  $\mathbf{s}^{l+1} = \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}$ 
10   $l \leftarrow l + 1$ 
11 end

```

---

*Remark:* in practical code implementations, the rank-1 data matrices  $\mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top$ ,  $i = 1, 2, \dots, n$ , involved in the computation of (56), are only necessary through the terms  $\mathcal{L}^*(\mathbf{x}_{i,*} \mathbf{x}_{i,*}^\top)$ , which can be readily pre-computed before the starting of the iterative process.

**Theorem 6** *Algorithm 4 subsequently converges for any sufficiently large  $\rho$ , that is, the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 4 has at least one limit point, and each limit point is a stationary point of (50).*

**Proof** The proof is deferred to Appendix C.4. ■

#### 4.7 $k$ -component heavy-tailed graphs

Extending Problem (50) for  $k$ -component graphs follows the same strategy as in Problem (42), which results in the following optimization program

$$\begin{aligned} & \underset{\mathbf{w} \geq 0, \Theta \succeq 0, \mathbf{V}}{\text{minimize}} \quad \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right) - \log \det^* (\Theta) + \eta \text{tr}(\mathcal{L} \mathbf{w} \mathbf{V} \mathbf{V}^\top), \\ & \text{subject to} \quad \Theta = \mathcal{L} \mathbf{w}, \quad \text{rank}(\Theta) = p - k, \quad \mathbf{d} \mathbf{w} = \mathbf{d}, \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \quad \mathbf{V} \in \mathbb{R}^{p \times k}. \end{aligned} \quad (57)$$

Following the exposition in the preceding sections, the partial augmented Lagrangian function of Problem (57) is given as

$$\begin{aligned} L_\rho(\Theta, \mathbf{w}, \mathbf{Y}, \mathbf{y}) &= \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right) - \log \det^* (\Theta) + \eta \text{tr}(\mathcal{L} \mathbf{w} \mathbf{V} \mathbf{V}^\top) \\ &+ \langle \mathbf{y}, \mathbf{d} \mathbf{w} - \mathbf{d} \rangle + \frac{\rho}{2} \|\mathbf{d} \mathbf{w} - \mathbf{d}\|_2^2 + \langle \mathbf{Y}, \Theta - \mathcal{L} \mathbf{w} \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L} \mathbf{w}\|_{\text{F}}^2. \end{aligned} \quad (58)$$

The subproblems for the variables  $\Theta$  and  $\mathbf{V}$  are identical to those of Problem (42), hence they follow the same update expressions.

The subproblem for  $\mathbf{w}$  is virtually the same as in (55), except for the additional term  $\eta \text{tr}(\mathcal{L} \mathbf{w} \mathbf{V}^l \mathbf{V}^{l\top}) = \eta \langle \mathbf{w}, \mathcal{L}^*(\mathbf{V}^l \mathbf{V}^{l\top}) \rangle$ . Hence, its update is also a projected gradient descent step, alike (35) where

$$\mathbf{a}^j \triangleq \mathcal{L}^* \left( \tilde{\mathbf{S}}^j + \eta \mathbf{V}^l \mathbf{V}^{l\top} - \mathbf{Y}^l - \rho \left( \Theta^{l+1} - \mathcal{L} \mathbf{w}^j \right) \right). \quad (59)$$

The dual variables  $\mathbf{Y}$  and  $\mathbf{y}$  are updated as in (36) and (37), respectively.

Algorithm 5, named ktGL, summarizes the implementation to solve Problem (57).

**Theorem 7** *Algorithm 5 subsequently converges for any sufficiently large  $\rho$ , that is, the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 5 has at least one limit point, and each limit point is a stationary point of (57).*

**Proof** See Appendix C.5. ■

## 5. Experimental Results

We perform experiments with price data queried from S&P500 stocks. In such real-world data experiments, where the a ground-truth graph cannot possibly be known, we evaluate the performance of the learned graphs by visualizing the resulting estimated graph network and verifying whether it is aligned with prior, expert knowledge available, *e.g.*, the GICS sector information of each stock<sup>4</sup>. In addition, we employ measures such as graph modularity (*cf.* Definition (68)) and density as an objective criterion to evaluate the quality of the estimated graphs.

4. It is important to notice that the GICS sector classification system might itself be prone to misclassifications specially for companies that serve many markets.

**Algorithm 5:**  $k$ -component Student- $t$  graph learning (ktGL)

---

**Data:** Data matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$ , initial estimate of the graph weights  $\mathbf{w}^0$ , desired number of graph components  $k$ , desired degree vector  $\mathbf{d}$ , degrees of freedom  $\nu$ , penalty parameter  $\rho > 0$ , tolerance  $\epsilon > 0$

**Result:** Laplacian estimation:  $\mathcal{L}\mathbf{w}^*$

```

1 initialize  $\mathbf{Y} = \mathbf{0}$ ,  $\mathbf{y} = \mathbf{0}$ 
2  $l \leftarrow 0$ 
3 while  $\max(|\mathbf{r}^l|) > \epsilon$  or  $\max(|\mathbf{s}^l|) > \epsilon$  do
4    $\triangleright$  update  $\Theta^{l+1}$  via (28)
5    $\triangleright$  update  $\mathbf{w}^{l+1}$  as in (35) with  $\mathbf{a}^j$  given as in (59)
6    $\triangleright$  update  $\mathbf{V}^{l+1}$  as in (48)
7    $\triangleright$  update  $\mathbf{Y}^{l+1}$  as in (36)
8    $\triangleright$  update  $\mathbf{y}^{l+1}$  as in (37)
9    $\triangleright$  compute residual  $\mathbf{r}^{l+1} = \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}$ 
10   $\triangleright$  compute residual  $\mathbf{s}^{l+1} = \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}$ 
11   $l \leftarrow l + 1$ 
12 end

```

---

*Baseline algorithms:* We compare the proposed algorithms (Table 1) with state-of-the-art, baseline algorithms, depending on the specific graph structure that they are suitable to estimate. In the existing literature, it is a common practice not to compare graph algorithms that adopt distinct operational assumptions, *i.e.*, the LGMRF approach and the smooth signal approach. This separation is certainly useful from a theoretical perspective. In this work, however, we are mostly interested in the applicability of graph learning algorithms in practical scenarios and whether the estimated graphs are aligned with prior expert knowledge available irrespective of their underlying assumptions. Therefore, in our experimental analysis, we consider algorithms from both operational assumptions. A summary of the baseline algorithms along with their target graph structure is illustrated in Table 2. For a fair comparison among algorithms, we set the degree vector  $\mathbf{d}$  equal to  $\mathbf{1}$  for the proposed algorithms, *i.e.*, we do not consider any prior information on the degree of nodes.

*Initial graph:* Because the algorithms proposed in this paper work in an iterative fashion, they naturally require an initial estimate of the graph. An appropriate initial estimation is critical to obtain a meaningful solution, especially in cases when the optimization problem is non-convex. However, obtaining an initial estimate inherently involves a trade-off between computational efficiency and quality. The latter being measured by how far the initial point is from an actual optimal point. Since the computational complexity of the proposed algorithms are bounded below by the eigenvalue decomposition  $O(p^3)$ , we are interested in simple strategies. Here we consider the strategy used by Kumar et al. (2019a), where the initial graph  $\mathbf{w}^0$  is set as  $(\tilde{\mathbf{w}})^+$ , where  $\tilde{\mathbf{w}}$  is the upper triangular part of the pseudo sample inverse correlation matrix  $\mathbf{S}^\dagger$ .

In the experiments that follow, we use daily price time series data of stocks belonging to the S&P500 index. We start by constructing the log-returns data matrix, *i.e.*, a matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$ , where  $n$  is the number of price observations and  $p$  is the number of stocks, such

Table 1: Proposed algorithms, their target graph structure, operational assumption, and computational complexity.

Algorithm	Graph Structure	Assumption	Complexity
tGL	connected	Laplacian Student- $t$ MRF	$O(p^3)$
kGL	$k$ -component	LGMRF	$O(p^3)$
ktGL	$k$ -component	Laplacian Student- $t$ MRF	$O(p^3)$

Table 2: Baseline algorithms, their target graph structure, operational assumption, and computational complexity.

Algorithm	Graph Structure	Assumption	Complexity
GL-SigRep (Dong et al., 2016)	connected	smooth signals	$O(np^2)$
SSGL (Kalofolias, 2016)	connected	smooth signals	$O(p^2)$
GLE-ADMM (Zhao et al., 2019)	connected	LGMRF	$O(p^3)$
NGL-MCP (Ying et al., 2020a)	connected	LGMRF	$O(p^3)$
SGL (Kumar et al., 2019a, 2020)	$k$ -component	LGMRF	$O(p^3)$
CLR (Nie et al., 2016)	$k$ -component	smooth signals	$O(p^3)$

that the  $j$ -th column contains the time series of log-returns of the  $j$ -th stock, which can be computed as

$$X_{i,j} = \log P_{i,j} - \log P_{i-1,j}, \quad (60)$$

where  $P_{i,j}$  is the closing price of the  $j$ -th stock on the  $i$ -th day.

### 5.1 $k$ -component graphs: degree control is crucial

To illustrate the importance of controlling the nodes degrees while learning  $k$ -component graphs, we conduct a comparison between the spectral constraints algorithm proposed in (Kumar et al., 2020), denoted as **SGL**, and the proposed  $k$ -component graph learning (Algorithm 3) on the basis of the sample correlation matrix. To that end, we set up experiments with two datasets: (i) stocks from four sectors, namely, **Health Care**, **Consumer Staples**, **Energy**, and **Financials**, from the period starting from Jan. 1st 2014 to Jan. 1st 2018. This datasets results in  $n = 1006$  stock price observations of  $p = 181$  stocks; (ii) we expand the dataset by including two more sectors, namely, **Industrials** and **Information Technology**. In addition, we collect data from Jan. 1st 2010 to Jan. 1st 2018, resulting in  $p = 292$  stocks and  $n = 2012$  observations.

Figure 5 shows the estimated financial stocks networks with  $k = 4$  (Figures 5a and 5b) and  $k = 6$  (Figures 5c and 5d). Clearly, the absence of degrees constraints in the learned graph by **SGL** (benchmark) (Kumar et al., 2020) shows evidence that the algorithm is unable to recover a non-trivial  $k$ -component solution, *i.e.*, a graph without isolated nodes. In addition, the learned graphs by **SGL** present a high number of inter-cluster connections (grey-colored edges), which is not expected from prior expert knowledge of the sectors. The proposed algorithm not only avoids isolated nodes via graph degree constraints, but most importantly learns graphs with meaningful representations, *i.e.*, they are aligned with the available sector information.

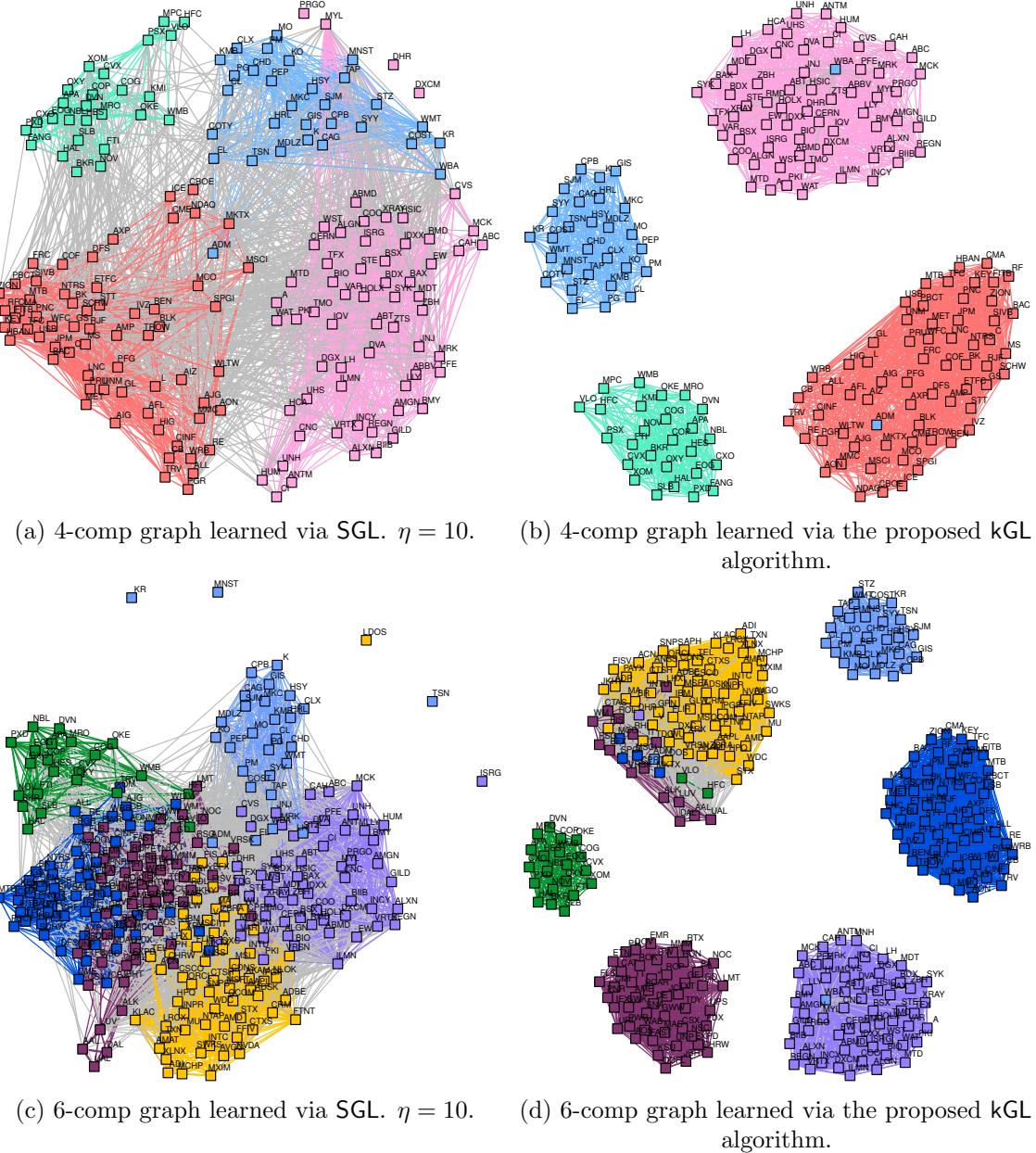


Figure 5: Rank constraints are met by the SGL algorithm (Kumar et al., 2020) (Figures 5a, 5c), nonetheless the learned graph conveys little information due to the lack of control on the degrees, which allows the learning of trivial  $k$ -component graphs, *i.e.*, those containing isolated nodes.

## 5.2 Effects of market factor and data preprocessing

Removing the market factor prior to performing analysis on a set of stock prices is a common practice (Mantegna, 1999; Laloux et al., 2000). The market factor is the component of stock signals associated with the strongest spectral coefficient. As we have argued in Section 3,

removing the market when learning graph matrices is implicitly done via the constraint  $\mathbf{L1} = \mathbf{0}$ , for the estimation of the Laplacian matrix, or via the construction of the  $\mathbf{Z}$  matrix for the estimation of the adjacency matrix. Therefore, it is not necessary to remove the market factor when learning graphs. Another guideline presented in Section 3 is that one should use the correlation matrix of the stock times series (or, equivalently, rescale the data such that each stock time series has unit variance) so as to obtain a meaningful cluster representation.

In order to verify these claims in practice, we set up an experiment where we collected price data from Jan. 3rd 2014 to Dec. 29th 2017 ( $n = 1006$  observations) of 82 selected stocks from three sectors: 28 from Utilities, 31 from Real Estate, and 23 from CommunicationServices.

We then proceed to learn four graphs, using the proposed kGL algorithm, with the following settings for the input data:

1. No data scaling and with market signal removed (Figure 6a).
2. No data scaling and with market signal included, *i.e.*, no data preprocessing (Figure 6b).
3. Scaled data and with market signal removed (Figure 6c).
4. Scaled data and with market signal included (Figure 6d).

The market signal is removed via eigenvalue decomposition of the sample correlation (covariance) matrix, where the largest eigenvalue is set to be zero.

Figures 6a and 6b depict evidence that using the sample covariance matrix (or equivalently, not scaling the input data), regardless of whether the market signal has been removed, leads to a graph with possibly many spurious connections (grey edges) that is not in agreement with the GICS sector classification. Figures 6c and 6d show that using the sample correlation matrix (or equivalently, scaling the stock time series such that they have the same variance) prior to learning the graph clearly shows meaningful graphical representations from stocks belonging to three distinct sectors, regardless of whether the market signal has been removed. In addition, the relative error between the estimated graphs in Figures 6a and 6b is 0.09, whereas the relative error between the estimated graphs in Figures 6c and 6d is  $4.6 \cdot 10^{-5}$ . Those relative error measurements further confirm that removing the market has little effect on the estimated graph due to the constraint  $\mathbf{L1} = \mathbf{0}$ , as explained in Section 3.

In addition, it has been argued that the sample correlation matrix may not always be a good measure of dependency for highly noisy, often non-linear dependent signals such as log-returns of stocks (de Prado, 2020). In the proposed framework, other measures of similarities can be used in place of the sample correlation matrix. For instance, we learn a graph under the same settings as the aforementioned experiment, but using the normalized mutual information,  $\bar{I}$ , between the log-return signals (assuming they follow a Gaussian distribution) as the input similarity matrix, which may be computed as

$$\bar{I}_{ij} = \begin{cases} -\frac{1}{2} \log(1 - \bar{S}_{ij}^2), & \text{if } i \neq j, \\ 1, & \text{else,} \end{cases} \quad (61)$$

where  $\bar{S}_{ij}^2$  is the sample correlation coefficient between the log-returns of stock  $i$  and  $j$ .

Figure 7 depicts the graph structure learned using the normalized mutual information. As it can be observed, the structure of Figure 7 is very similar to that of Figure 6c. Objectively,

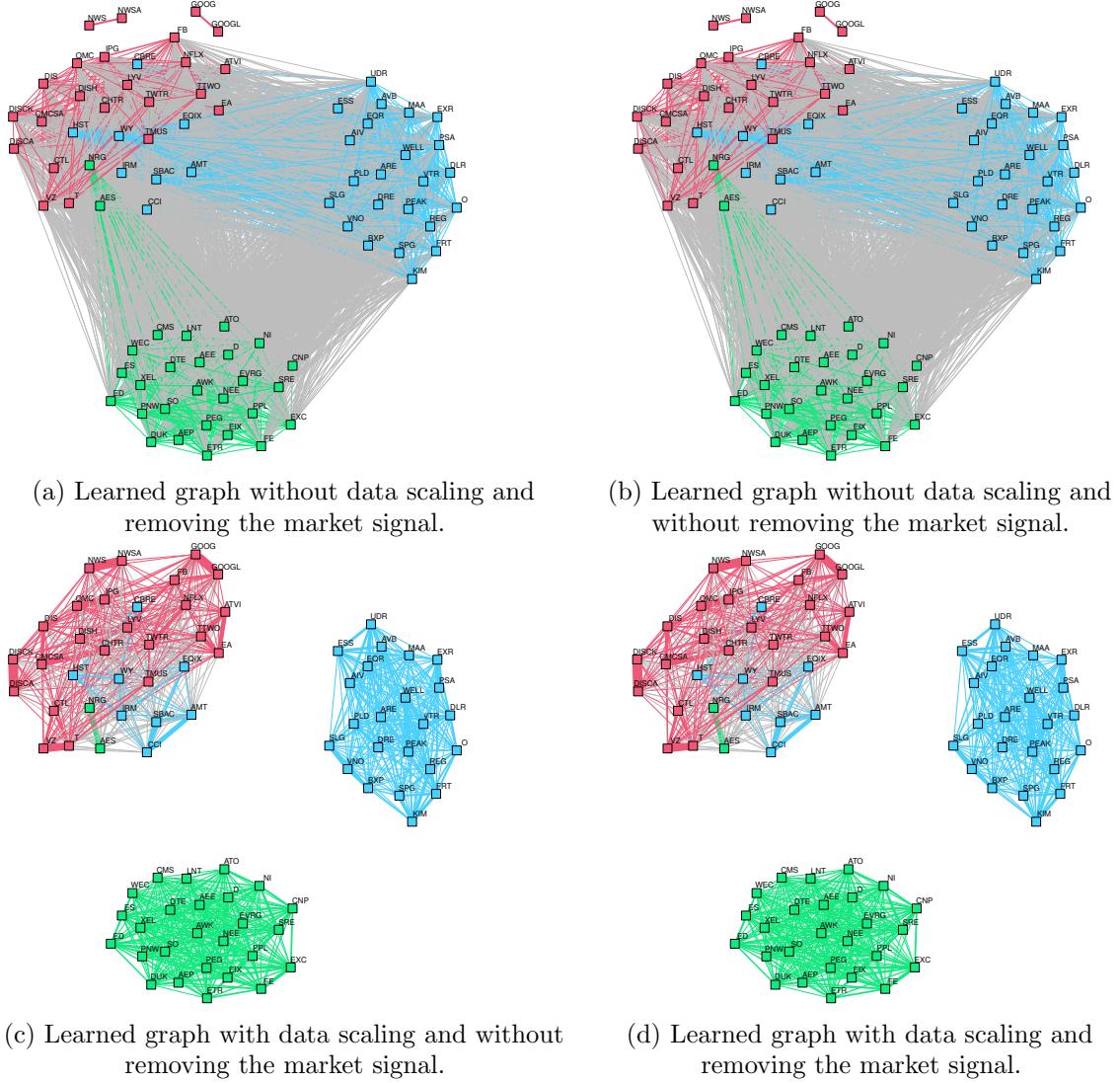


Figure 6: Effects of data preprocessing on the learned graphs.

the f-score between the learned graphs is 0.91 while the relative error is 0.35, which may indicate that using either the normalized mutual information or the sample correlation matrix are equally acceptable inputs for the learning algorithm.

Finally, we use the state-of-the-art, two-stage CLR algorithm (Nie et al., 2016) to learn a 3-component graph for the selected stocks on the basis of the scaled input data matrix. Figure 8 depicts the learned graph network. As it can be observed, unlike the proposed algorithm, CLR clusters together most of the stocks belonging to the Real State and Communication Services sectors, which is not expected from an expert *prior* information such as GICS.

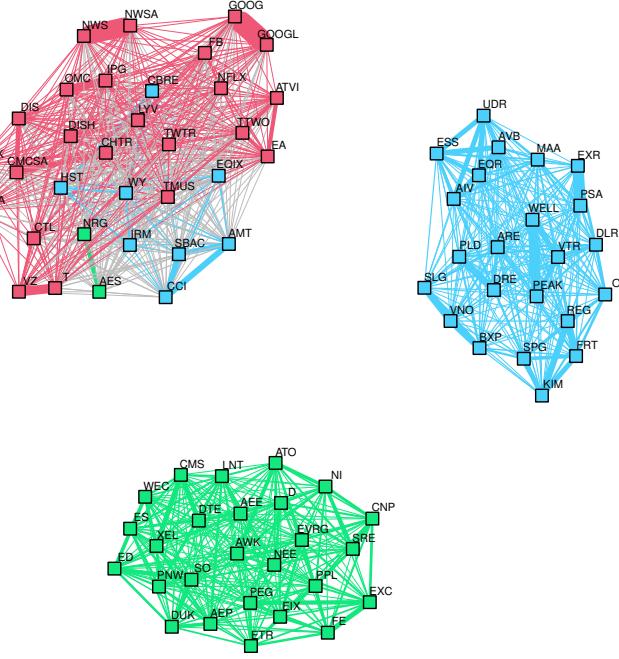


Figure 7: Learned graph with the proposed kGL algorithm 3 using the normalized mutual information as input matrix.

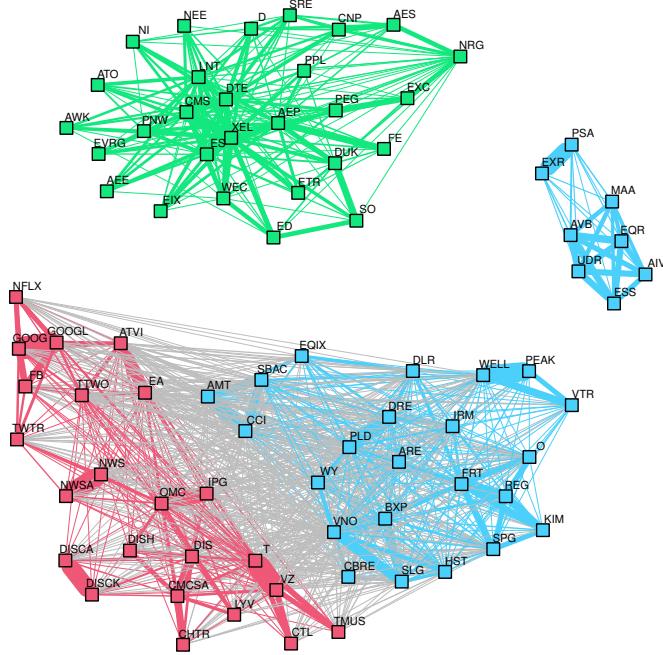


Figure 8: Learned graph with the CLR algorithm (Nie et al., 2016) with unit-variance, scaled log-return data matrix as input.

### 5.3 Heavy-tails effects: warm-up

In this experiment, we would like to convey the advantages of using graph learning algorithms based on the assumptions that the data is heavy-tailed. To that extent, we compare three Laplacian-constrained models: (1) Gaussian (GLE-ADMM ([Zhao et al., 2019](#))) (2) Gaussian with minimax concave sparsity penalty (NGL-MCP ([Ying et al., 2020b](#))), and (3) Student- $t$  ([tGL Algorithm 4](#)). These models are investigated under two scenarios: (i) strong ( $\nu \approx 4$ ) and (ii) weak ( $\nu \approx 10$ ) presence of heavy-tails. On both scenarios, we selected stocks belonging to five different sectors, namely: Consumer Staples, Consumer Discretionary, Industrials, Energy, and Information Technology.

*Remark on hyperparameters:* For the Gaussian model with minimax concave penalty, we tune the sparsity hyperparameter so that the estimated graph obtains the highest modularity. While tuning an one-dimensional hyperparameter may not pose issues while performing post-event analysis, it does compromise the performance of real-world online systems where the value of such hyperparameter is often unknown and data-dependent. For the Student- $t$  model, the degrees of freedom  $\nu$  can be computed in a prior stage directly from the data using, *e.g.*, the methods in ([Liu et al., 2019](#)), or in a sliding-window fashion for the case of real-time systems.

**Strong heavy-tails:** for this experiment, we queried data from 222 stocks from Jan. 3rd 2008 to Dec. 31st 2009, which represents 504 data observations per stock, resulting in a sample-parameter size ratio of  $n/p \approx 2.27$ . This particular time-frame presents a high amount of volatility due to the 2008 US depression. To quantify the extent of heavy-tails in the data, we fit a multivariate Student- $t$  distribution using the matrix of log-returns  $\mathbf{X}$ , where we obtain  $\nu \approx 4.06$ , which indeed indicates a high presence of heavy-tailed data points. In addition, we measured the average annualized volatility across all stocks and obtained  $\text{volatility} \approx 0.53$ . Figure 9 provides a summary of this market scenario.

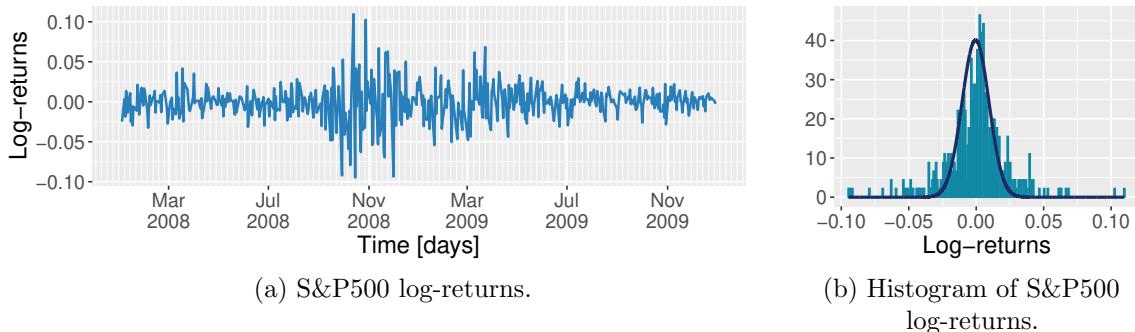


Figure 9: State of the US stock market, as captured by the S&P500 index, on the **strong heavy-tails** scenario, which starts from Jan. 3rd 2008 until Dec. 31st 2009. Figure 9a shows the S&P500 log-returns time series, where the increase in volatility due to the global financial crisis in 2008 is clearly noticeable. Figure 9b shows a histogram of the S&P500 log-returns during the aforementioned time period, where the solid curve represents a Gaussian fit. It can be noticed that the tails of the Gaussian decays much faster than the tails of the empirical histogram, indicating the presence of heavy-tails or outliers.

**Weak heavy-tails:** in this scenario, we collected data from 204 stocks from Jan. 5th 2004 to Dec. 30th 2006, which represents 503 data points per stock, resulting in a sample-parameter size ratio of  $n/p \approx 2.47$ . During this time-window, the market was operating relatively nominal. By fitting a multivariate Student- $t$  distribution to the matrix of log-returns, we obtain  $\nu \approx 10.11$ , which indicates little presence of outliers, and that the data is nearly Gaussian. The average annualized volatility measured across all stocks is  $\text{volatility} \approx 0.27$ , which is half of the annualized volatility in the strong heavy-tails case. Figure 10 provides a summary of this market scenario.

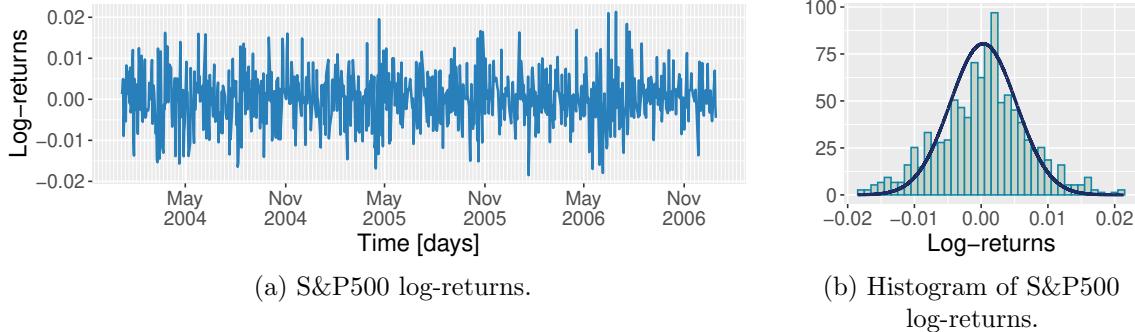


Figure 10: State of the US stock market, as captured by the S&P500 index, on the **weak heavy-tails** scenario, which starts from Jan. 5th 2004 until Dec. 30th 2006. Figure 10a shows the S&P500 log-returns time series, where no noticeable volatility clustering event is present, while Figure 10 depicts its histogram along with a Gaussian fit that closely matches the empirical distribution.

Figure 11 depicts the learned stock graphs on these scenarios. In either scenario, it can be readily noticed that the graphs learned with the Student- $t$  distribution are sparser than those learned with the Gaussian assumption, which results from the fact that the Gaussian distribution is more sensitive to outliers. As for the Gaussian graphs with sparsity, they present a significant improvement when compared to the non-sparse counterpart. The Student- $t$  graphs, on the other hand, present the highest degree of interpretability as measured by their higher modularity value and ratio between the number of intra-sector edges and inter-sector edges (*cf.* Tables 3 and 4), which is the expected behavior from stock sector classification systems such as GICS.

Among the learned Gaussian graphs (Figures 11a and 11d), it can be seen that the learned graph in the weak heavy-tailed scenario presents a cleaner graphical representation, by having less inter-sector and intra-sector edges, while also having a higher graph modularity (*cf.* Tables 3 and 4), than that of the Gaussian graph in the strong heavy-tailed scenario.

Table 3: Edge distribution for the **strong** heavy-tails case.

model	inter-sector edges	intra-sector edges	modularity ( $Q$ )
Gaussian	2918	2615	0.23
Gaussian w/ sparsity	158	339	0.46
Student- $t$	137	384	<b>0.51</b>

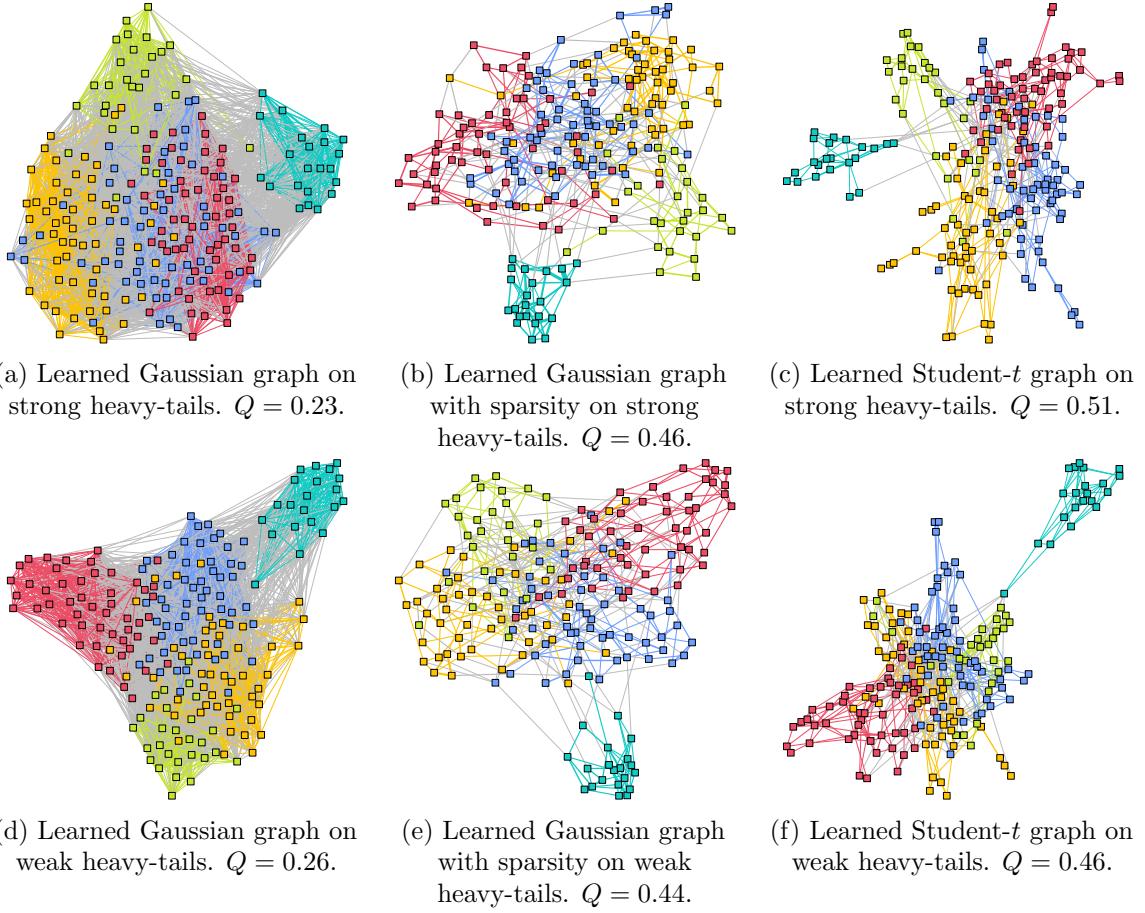


Figure 11: Learned graph networks with Gaussian (Figures 11a and 11d), Gaussian with sparsity (Figures 11b and 11e), and Student- $t$  (Figures 11c and 11f), for contrasting heavy-tail scenarios.

Table 4: Edge distribution for the **weak** heavy-tails case.

model	inter-sector edges	intra-sector edges	modularity ( $Q$ )
Gaussian	2028	1966	0.26
Gaussian w/ sparsity	173	325	0.44
Student- $t$	197	438	<b>0.46</b>

#### 5.4 Heavy-tails effects: additional analysis

In this section, we perform a similar analysis as in the previous experiment, except that we consider stocks from the sectors **Industrials**, **Consumer Staples**, **Consumer Discretionary**, **Information Technology**, **Energy**, **Health Care**, and **Real State**.

**Strong heavy-tails:** for this experiment, we queried data from 347 stocks from Jan. 5th 2016 to Dec. 23rd 2020, which represents 1253 data observations per stock, resulting in a sample-parameter ratio of  $n/p \approx 3.61$ . This particular time-frame presents an extreme

high amount of volatility around the beginning of 2020 due to the financial crisis caused by the COVID-19 pandemic. To quantify the amount of outliers in this time frame, we fit a multivariate Student- $t$  distribution using the matrix of log-returns  $\mathbf{X}$ , where we obtain  $\nu \approx 4.15$ , which indeed indicates a high presence of heavy-tailed data points. In addition, we measured the average annualized volatility across all stocks and obtained volatility  $\approx 0.34$ . Figure 12 provides a summary of this market scenario.

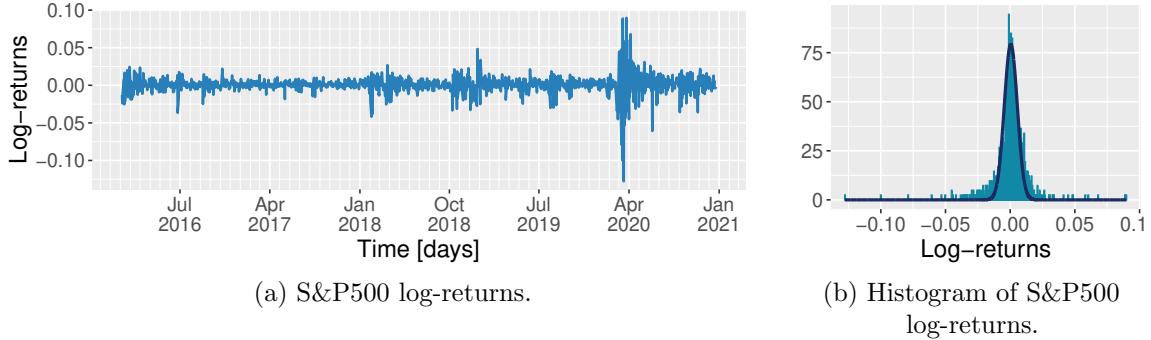


Figure 12: State of the US stock market, as captured by the S&P500 index, on the **strong heavy-tails** scenario, which starts from Jan. 5th 2016 until Jul. 20th 2020. Figure 12a shows the S&P500 log-returns time series, where the increase in volatility due to the COVID-19 pandemic is prominent. Figure 12b illustrates the empirical distribution of the S&P500 log-returns along with its Gaussian fit. It can be noticed that events far beyond the tails decay are present.

**Moderate heavy-tails:** for this setting, we queried data from 332 stocks from Jan. 2nd 2013 to Jun. 29th 2018, which represents 1383 data observations per stock, resulting in a sample-parameter ratio of  $n/p \approx 4.17$ . We fit a multivariate Student- $t$  distribution using the matrix of log-returns  $\mathbf{X}$ , where we obtain  $\nu \approx 7.11$ . In addition, we measured the annualized average volatility across all stocks and obtained volatility  $\approx 0.25$ . Figure 13 provides a summary of this market scenario.

Figure 14 depicts the learned graphs on the aforementioned scenarios. Similarly from the previous experiment, it can be noticed that in either scenario the graphs learned with Student- $t$  are indeed sparser, more modular, and hence, more interpretable, than those learned with the Gaussian assumption (*cf.* Tables 5 and 6).

The learned Gaussian graphs (Figures 14a and 14d), are very dense and present a high number of spurious connections (grey edges) among stocks that are arguably not related in practice. However, it can be noticed that the learned graph in the heavy-tailed scenario presents a cleaner graphical representation, by having less inter-sector edges, while also having a higher graph modularity (*cf.* Tables 6 and 5) than that of the Gaussian graph in the strong heavy-tailed scenario, which is consistent with the previous experiment. The usage of sparsity in does improve the Gaussian graphs (Figures 14b and 14e), but only to limited extent when compared to the graphs learned using the Student- $t$  assumption.

The Student- $t$  graphs, on the other hand, present a high degree of modularity, where most of the sectors can easily be identified. We also measured that the Student- $t$  distribution outputs graphs (Figure 14c and 14f) with higher graph modularity.

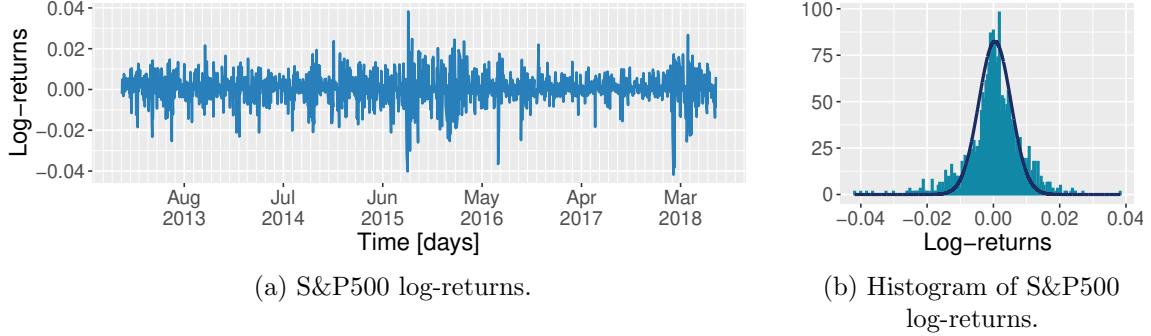


Figure 13: State of the US stock market, as captured by the S&P500 index, on the **moderate heavy-tails** scenario, which starts from Jan. 2nd 2013 until Jun. 29th 2018. Figure 13a shows the S&P500 log-returns time series, where a few significant heavy-tailed observations are noticeable, along with its histogram (Figure 13b) whose Gaussian fit indicates that the presence of outlier data points are mostly concentrated around the turning points of the tails.

Table 5: Edge distribution for the **strong** heavy-tails case.

model	inter-sector edges	intra-sector edges	modularity ( $Q$ )
Gaussian	4262	3947	0.31
Gaussian w/ sparsity	405	986	0.54
Student- $t$	124	579	<b>0.66</b>

Table 6: Edge distribution for the **moderate** heavy-tails case.

model	inter-sector edges	intra-sector edges	modularity ( $Q$ )
Gaussian	4198	4063	0.32
Gaussian w/ sparsity	375	1021	0.57
Student- $t$	131	627	<b>0.66</b>

## 5.5 Heavy-tails and $k$ -component graphs

In order to verify the learning of heavy-tail and  $k$ -component graphs jointly, we estimate a graphs of stocks using the datasets described in subsections 5.1 5.2 via the ktGL algorithm. Figure 15 depicts the learned graphs where we can observe sparse characteristics that agree with the connected graphs estimated in the preceding section. In addition, when compared to the Gaussian case (Figures 5b and 6c), the graphs estimated in Figures 15 and 15b reveal a finer, possibly more accurate description of the actual underlying stock market scenario.

## 5.6 Effect of crisis on the learned graphs: COVID-19 case of study

In the experiments that follow, we focus on illustrating the impact of the COVID-19 economic crisis on the learned graphs from the S&P500 stock market and the foreign exchange market. The COVID-19 pandemic affected the US stock market quite significantly especially throughout the month of March 2020.

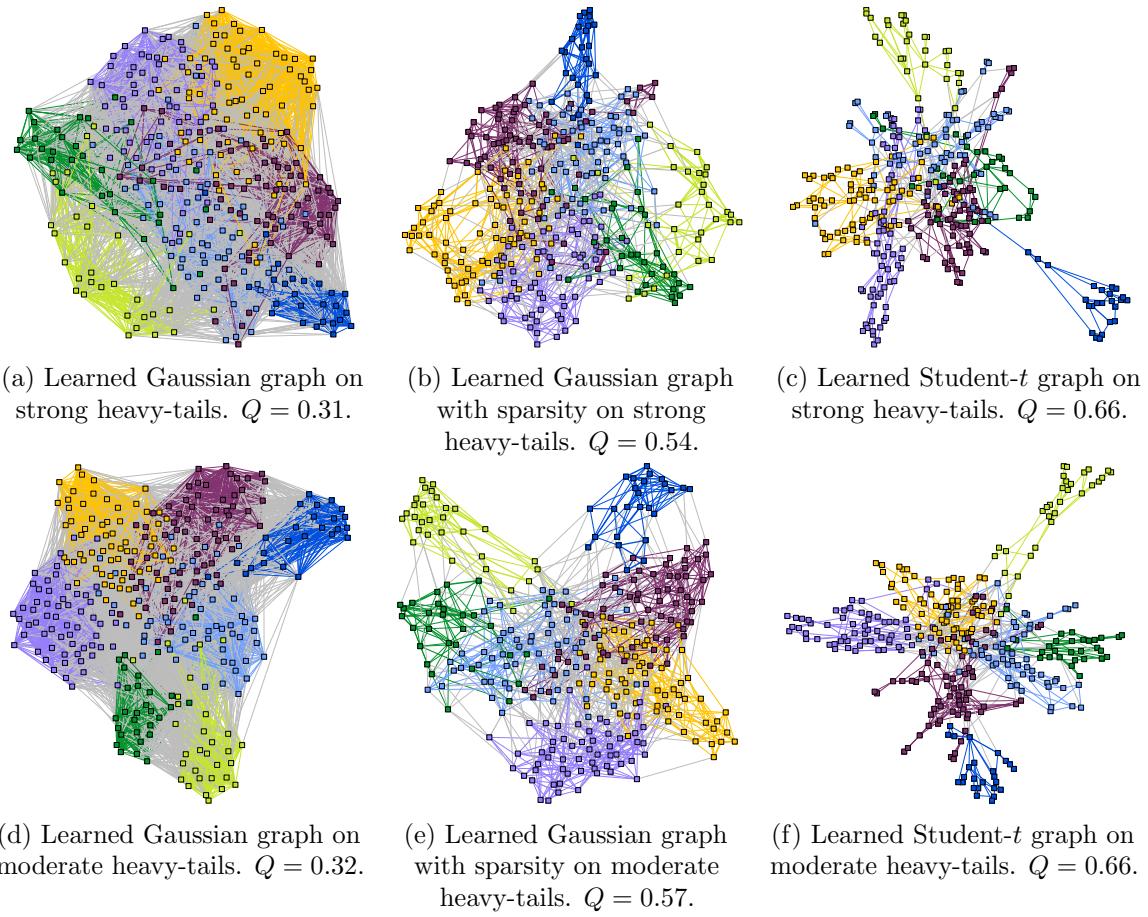


Figure 14: Learned graph networks with Gaussian (Figures 14a and 14d), Gaussian with sparsity (Figures 14b and 14e), and Student- $t$  (Figures 14c and 14f), for contrasting heavy-tail scenarios.

### 5.6.1 STOCKS

In this experiment, we investigate the effects of the financial crisis caused by the COVID-19 pandemic on the learned graphs of stocks.

We start by selecting 97 stocks across all 11 sectors of the S&P500 and computing their log-returns during two time frames: (i) from Apr. 22nd 2019 to Dec. 31st 2019 and (ii) from Jan. 2nd 2020 to Jul. 31st 2020.

Out of those stocks, nine of them showed growth in returns over the period of 24 days starting from Feb. 18th 2020 to March 20th 2020. Their symbols along with their monthly return during this period is summarized in Table 7. Figure 16 shows the log-returns of the selected stocks over the considered time period. In Figure 16b, the economic crisis is noticeable from the increase in the spread of the log-returns, throughout the month of March 2020, caused by the COVID-19 pandemic.

Figure 17 shows the learned networks using the proposed tGL algorithm 4 with Student- $t$  assumption. Figure 17a shows that the stocks with positive average linear return (blue

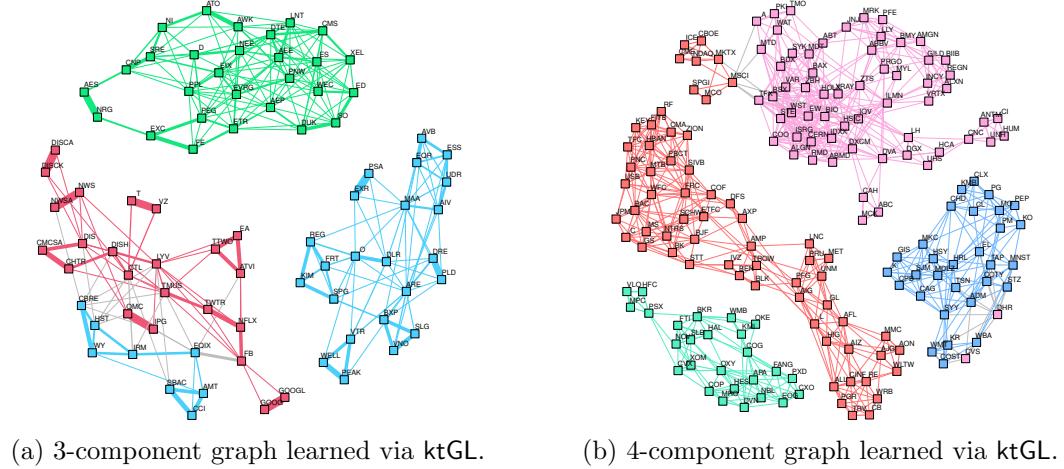


Figure 15: Learned graphs from different stock market scenarios taking into account both  $k$ -component requirements and heavy-tail assumptions.

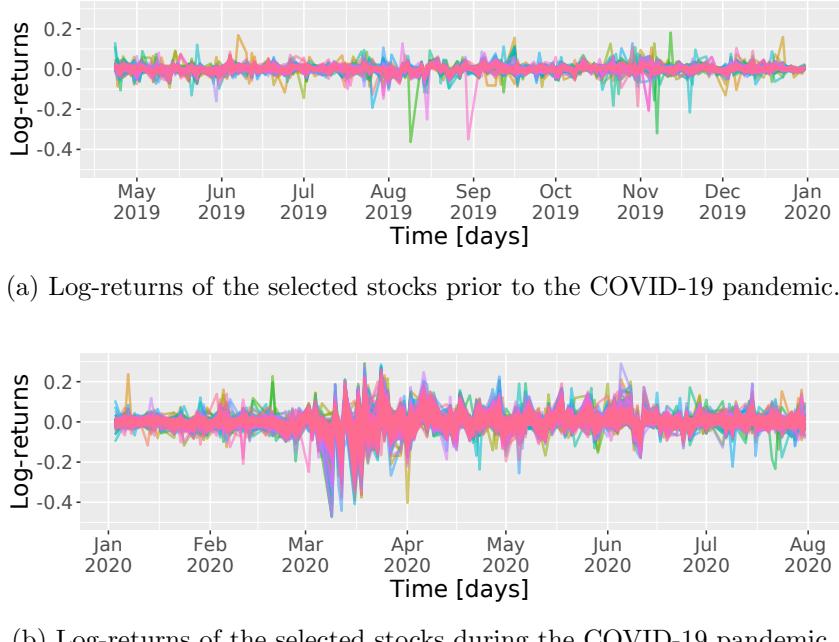


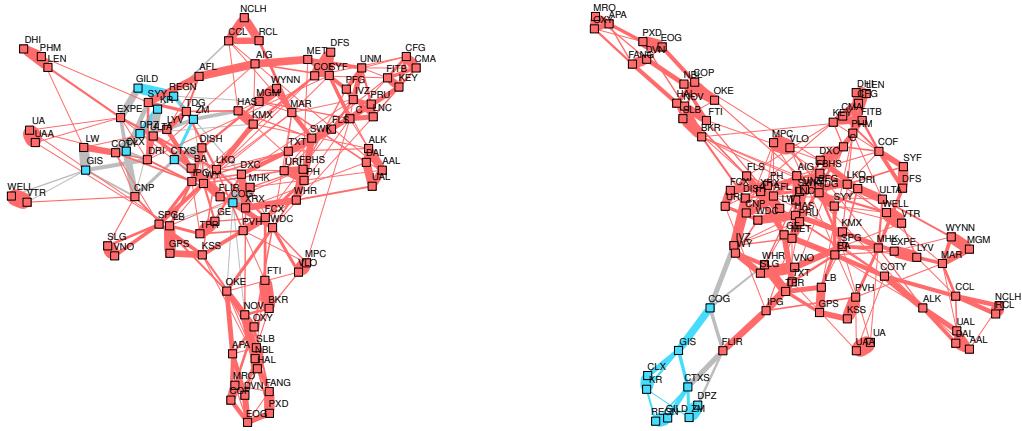
Figure 16: Log-returns of 97 selected stocks from April 22nd 2019 to July 31st 2020. The vertical axis is fixed on panels in order to better illustrate the change in volatility as a result of the COVID-19 pandemic.

squares) during the COVID-19 pandemic are not particularly correlated on the period before the pandemic. Figure 17b, on the other hand, shows that the learned graph is able to correctly cluster those stocks. In particular, it can be observed that the network of Figure 17b is objectively more modular than that of Figure 17a. This experiment shows

Table 7: Stocks with positive monthly return from 2020-02-15 to 2020-03-20.

Symbol	GICS Sector	monthly return
CLX	Consumer Staples	4.8%
COG	Energy	2.5%
CTXS	Information Technology	0.4%
DPZ	Consumer Discretionary	3.9%
GILD	Health Care	5.8%
GIS	Consumer Staples	0.9%
KR	Consumer Staples	6.9%
REGN	Health Care	6.1%
ZM	—	19.4%

evidence that the graph models can be employed as a tool to identify events of interest in the network.



(a) Learned Student- $t$  graph from data comprising the time window from Apr. 22nd 2019 to Dec. 31st 2019.  $Q = 0.024$ .

(b) Learned Student- $t$  graph from stock data during the financial crisis caused by COVID-19 from Jan. 2nd 2020 to Jul. 31st 2020.  $Q = 0.10$ .

Figure 17: Graphs of stocks learned with data prior and during the financial crisis caused by the COVID-19 pandemic in 2020. Figure 17a shows that before the COVID-19 pandemic the nodes in blue are somewhat independent among themselves. Figure 17b shows that during the COVID-19 pandemic the stocks highlighted in blue are strongly connected and separated from the rest of the graph, as in fact they showed a positive monthly return during the severe economic period between Feb. 18th and Mar. 20th, 2020.

In addition, we compare the proposed learned graphs in Figure 17 with graphs learned from algorithms that employ the smooth signal approach. Figure 18 shows the learned graphs from SSGL (Kalofolias, 2016) and GL-SigRep (Dong et al., 2016). As we can observe from both Figure 18a and 18b, the learned networks do not present a meaningful graph representation in this setting. While the clustering property of the stocks with positive monthly return is preserved in the network learned with GL-SigRep, it does not capture the

fine dependencies between pairs of stocks like the ones shown by the proposed kGL algorithm in Figure 17. In addition, tuning the hyperparameters in the smooth signal algorithms is an involved task.

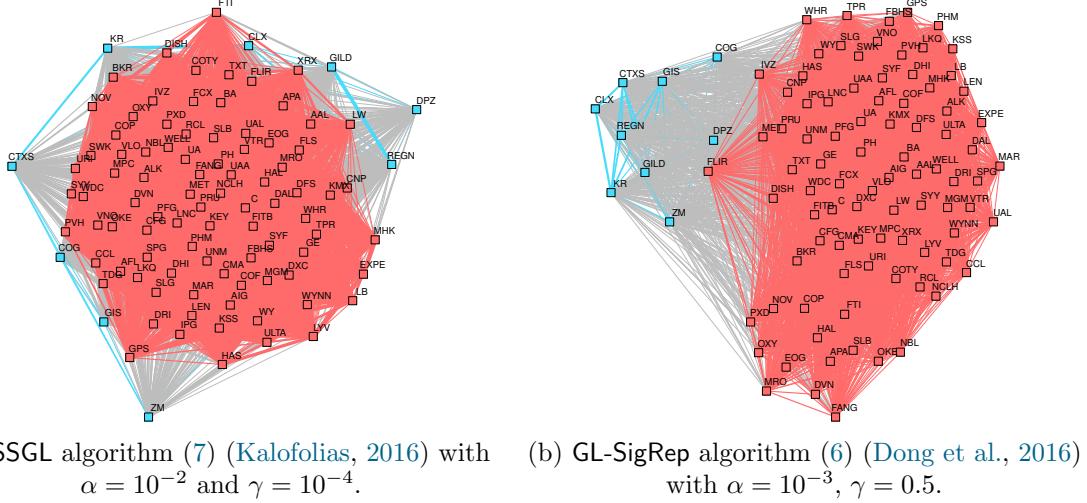


Figure 18: Learned graphs with existing smooth signal-based algorithms from stock data during the financial crisis caused by COVID-19.

### 5.6.2 FOREIGN EXCHANGE

We set up an experiment with data from the foreign exchange (FX) market, where, similarly to the previous experiment, we would like to investigate whether the learned graph is able to identify currencies that became more valuable with respect to the US dollar during the COVID-19 pandemic. To that end, we query FX data of the 34 most traded currencies as of 2019 in two time windows: (i) from Feb. 1st 2019 to May 1st 2019 and (ii) from Feb. 3rd 2020 to May 1st 2020.

We then obtain the list of currencies for which the US dollar became less valuable during the period from Feb. 15th to Apr. 15th 2020. Table 8 shows the list of such currencies along with the annualized return of the ratio USD/CUR, where CUR is a given currency.

Table 8: Currencies for which the US dollar had negative annualized return from 2020-02-15 to 2020-04-15.

Symbol	Name	Annualized return
EUR	Euro	-7.2%
JPY	Japanese Yen	-13.4%
CHF	Swiss Franc	-10.8%
HKD	Hong Kong Dollar	-1.1%
DKK	Danish Krone	-8.0%

We then use the proposed heavy-tail graph learning framework (tGL Algorithm 4) to learn the graph networks of foreign exchange data for the two aforementioned time frames.

Figure 19 depicts the learned graphs, where in Figure 19b clearly shows that the currencies that had an increase in value during the pandemic are clustered together (red edges), except for the HKD, which was the currency that presented the smallest increase. On the other hand, prior to the pandemic, the FX market behaved in a somewhat random fashion, which is seen by the smaller value in graph modularity.

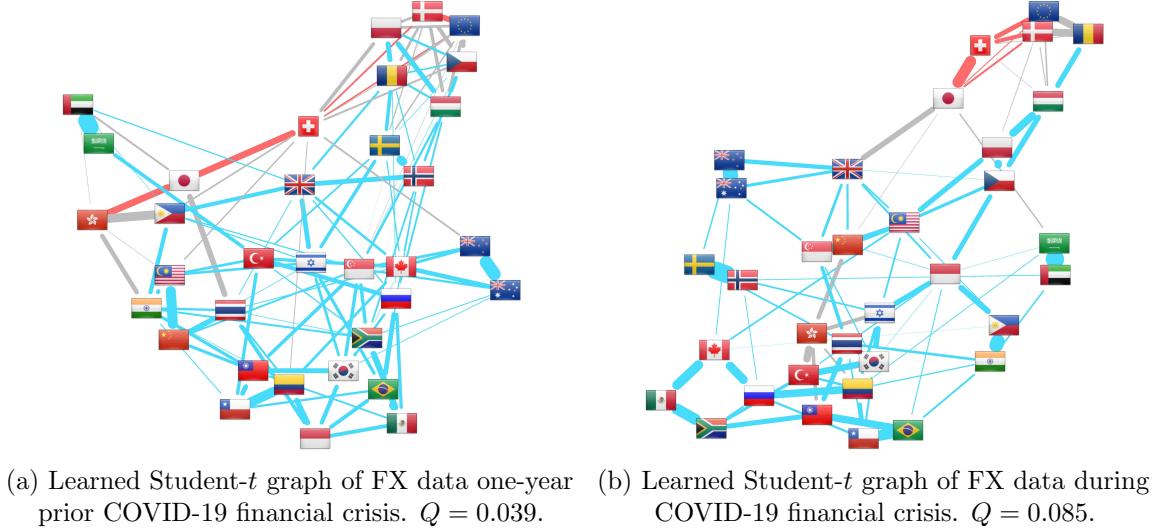


Figure 19: Learned graphs from FX data.

## 6. Conclusions

This paper has presented novel interpretations for Laplacian constraints of graphs from the perspective of financial data. Those interpretations serve as guidelines for users when it comes to apply graph learning algorithms to estimate networks of financial instruments such as stocks and currencies. We have also proposed novel algorithms based on the ADMM and MM frameworks for learning graphs from data. Those algorithms fill major gaps in the literature, especially on what concerns learning heavy-tailed and  $k$ -component graphs. In particular, the heavy-tail graph learning framework is paramount for financial data, which exceptionally outperforms conventional state-of-the-art algorithms derived on the assumption that the input data is Gaussian. Another feature of heavy-tail graphs is that they are naturally sparse. State-of-the-art sparse graph learning frameworks, while useful in many contexts beyond finance, are cumbersome to tune due to their hyperparameters. We, on the other hand, advocate that sparsity provided from heavy-tail distributions is a more principled way to obtain interpretable graph representations. In the case of  $k$ -component graphs, we proposed a principled, versatile framework that avoids isolated nodes via a simple linear constraint on the degree of the nodes. This extension allows, for instance, the estimation of particular types of graphs such as regular graphs. Moreover, the proposed graph algorithms have shown significant potential to capture nuances in the data caused by, *e.g.*, a financial crisis event. Finally, it is worth noting that the methods developed in this paper may be applicable to scenarios beyond financial markets, in particular, we envision

benefits for practical applications where the data distributions significantly departs from that of Gaussian.

## Acknowledgments

The numerical algorithms proposed in this work were implemented in the R language and made use of softwares such as **CVXR** (Fu et al., 2020), **Rcpp** (Eddelbuettel and Francois, 2011), **RcppEigen** (D. Bates, 2013), **RcppArmadillo** (Eddelbuettel and Sanderson, 2014), and **igraph** (Csárdi, 2019). This work was supported by the Hong Kong GRF 16207019 research grant.

## Appendix A. Empirical Convergence

In this supplementary section, we illustrate the empirical convergence performance of the proposed algorithms for the experimental settings considered. All the experiments were carried out in a MacBook Pro 13in. 2019 with Intel Core i7 2.8GHz, 16GB of RAM.

The quantities  $\mathbf{r}^l$  and  $\mathbf{s}^l$ , which are defined as  $\mathbf{r}^l = \Theta^l - \mathcal{L}\mathbf{w}^l$ ,  $\mathbf{s}^l = \mathbf{d}\mathbf{w}^l - \mathbf{d}$ , are the primal residuals and  $\mathbf{v}^l = \rho\mathcal{L}^*(\Theta^l - \Theta^{l-1})$  is the dual residual.

From Figures 20–25, we can observe that the norm of the residuals quantities quickly approach zero after a transient phase typical of ADMM-like algorithms.

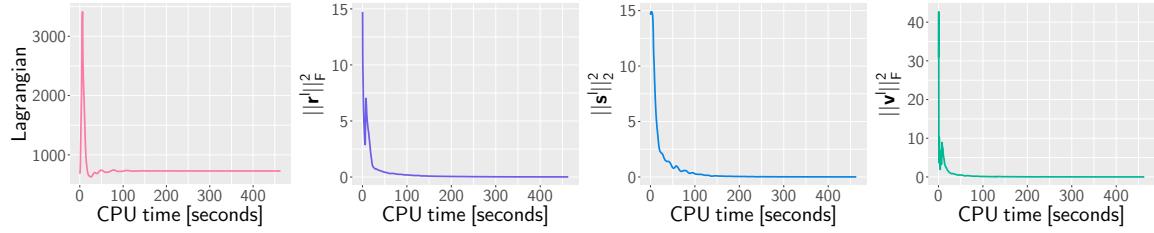


Figure 20: Empirical convergence for “warm-up” heavy-tail experiment data with Student- $t$  model.

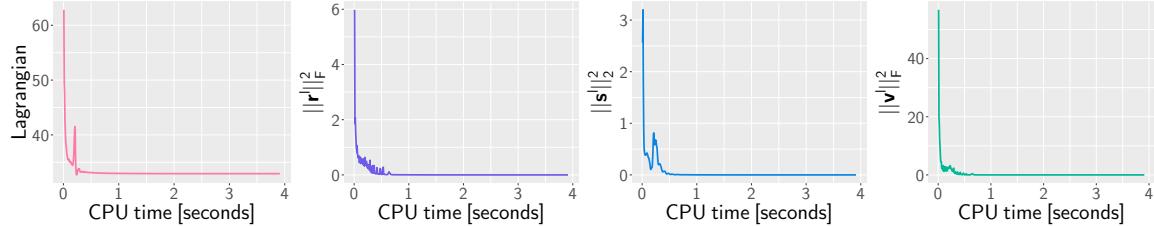


Figure 21: Empirical convergence for stock clustering with three sectors.

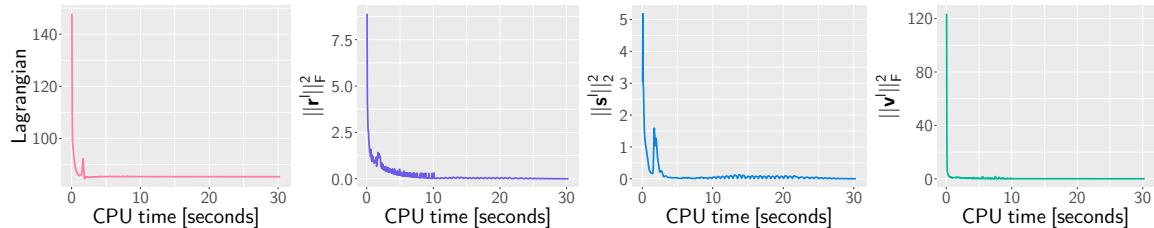


Figure 22: Empirical convergence for stock clustering with four sectors.

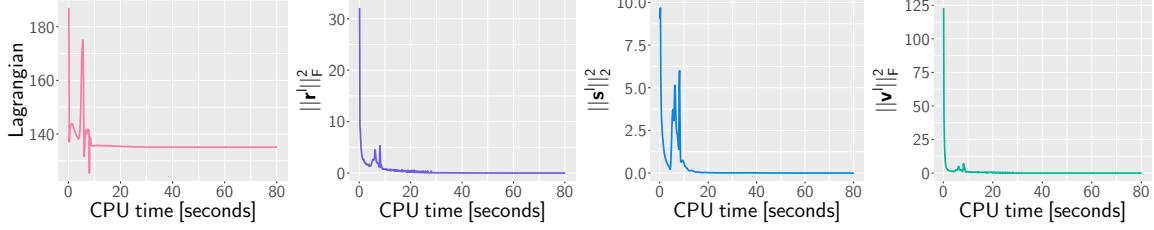


Figure 23: Empirical convergence for stock clustering with six sectors.

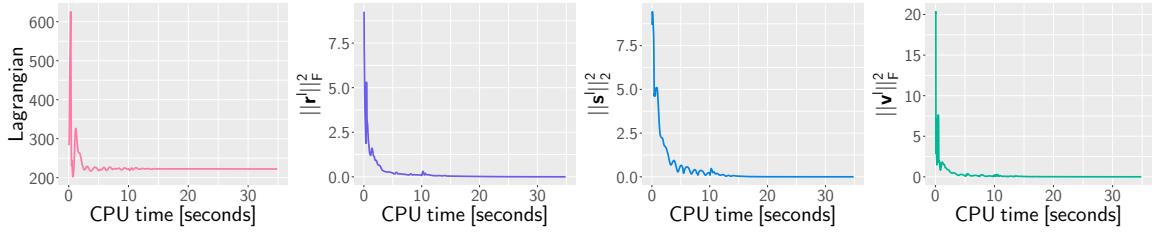


Figure 24: Empirical convergence for COVID-19 data experiment.

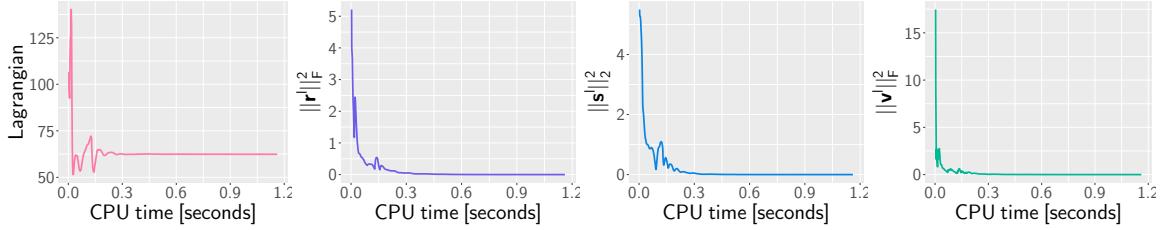


Figure 25: Empirical convergence for FX data experiment.

## Appendix B. Definitions

**Definition 8 (Laplacian operator)** The Laplacian operator (Kumar et al., 2019a)  $\mathcal{L} : \mathbb{R}_+^{p(p-1)/2} \rightarrow \mathbb{R}^{p \times p}$ , which takes a nonnegative vector  $\mathbf{w}$  and outputs a Laplacian matrix  $\mathcal{L}\mathbf{w}$ , is defined as

$$[\mathcal{L}\mathbf{w}]_{ij} = \begin{cases} -w_{i+s(j)}, & \text{if } i > j, \\ [\mathcal{L}\mathbf{w}]_{ji}, & \text{if } i < j, \\ -\sum_{i \neq j} [\mathcal{L}\mathbf{w}]_{ij}, & \text{if } i = j, \end{cases} \quad (62)$$

where  $s(j) = \frac{j-1}{2}(2p-j) - j$ .

**Definition 9 (Adjacency operator)** The adjacency operator (Kumar et al., 2020)  $\mathcal{A} : \mathbb{R}_+^{p(p-1)/2} \rightarrow \mathbb{R}^p$ , which takes a nonnegative vector  $\mathbf{w}$  and outputs an Adjacency matrix  $\mathcal{A}\mathbf{w}$ , is defined as

$$[\mathcal{A}\mathbf{w}]_{ij} = \begin{cases} w_{i+s(j)}, & \text{if } i > j, \\ [\mathcal{A}\mathbf{w}]_{ji}, & \text{if } i < j, \\ 0, & \text{if } i = j, \end{cases} \quad (63)$$

where  $s(j) = \frac{j-1}{2}(2p-j) - j$ .

**Definition 10 (Degree operator)** The degree operator  $\mathfrak{d} : \mathbb{R}^{p(p-1)/2} \rightarrow \mathbb{R}^p$ , which takes a nonnegative vector  $\mathbf{w}$  and outputs the diagonal of a Degree matrix, is defined as

$$\mathfrak{d}\mathbf{w} = (\mathcal{A}\mathbf{w}) \mathbf{1}. \quad (64)$$

**Definition 11 (Adjoint of Laplacian operator)** The adjoint of Laplacian operator (Kumar et al., 2019a)  $\mathcal{L}^* : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p(p-1)/2}$  is defined as

$$(\mathcal{L}^* \mathbf{P})_{s(i,j)} = \mathbf{P}_{i,i} - \mathbf{P}_{i,j} - \mathbf{P}_{j,i} + \mathbf{P}_{j,j} \quad (65)$$

where  $s(i,j) = i - j + \frac{j-1}{2}(2p-j)$ ,  $i > j$ .

**Definition 12 (Adjoint of adjacency operator)** The adjoint of adjacency operator (Kumar et al., 2019a)  $\mathcal{A}^* : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p(p-1)/2}$  is defined as

$$(\mathcal{A}^* \mathbf{P})_{s(i,j)} = \mathbf{P}_{i,i} + \mathbf{P}_{j,j} \quad (66)$$

where  $s(i,j) = i - j + \frac{j-1}{2}(2p-j)$ ,  $i > j$ .

**Definition 13 (Adjoint of degree operator)** The adjoint of degree operator  $\mathfrak{d}^* : \mathbb{R}^p \rightarrow \mathbb{R}^{p(p-1)/2}$  is given as

$$(\mathfrak{d}^* \mathbf{y})_{s(i,j)} = \mathbf{y}_i + \mathbf{y}_j, \quad (67)$$

where  $s(i,j) = i - j + \frac{j-1}{2}(2p-j)$ ,  $i > j$ .

An alternative expression for  $\mathfrak{d}^*$  is  $\mathfrak{d}^* \mathbf{y} = \mathcal{L}^* \text{Diag}(\mathbf{y})$ , where  $\mathcal{L}^*$  is the adjoint of the Laplacian operator.

**Definition 14 (Modularity)** The modularity of a graph (Newman, 2006) is defined as  $Q : \mathbb{R}^{p \times p} \rightarrow [-1/2, 1]$ :

$$Q(\mathbf{W}) \triangleq \frac{1}{p(p-1)} \sum_{i,j} \left( \mathbf{W}_{ij} - \frac{d_i d_j}{p(p-1)} \right) \mathbb{1}(t_i = t_j), \quad (68)$$

where  $d_i$  is the weighted degree of the  $i$ -th node, i.e.  $d_i \triangleq [\mathfrak{d}(\mathbf{w})]_i$ ,  $t_i \triangleq f_t(i)$ ,  $i \in \mathcal{V}$ , is the type of the  $i$ -th node, and  $\mathbb{1}(\cdot)$  is the indicator function.

**Definition 15 (Proximal Operator)** The proximal operator of a function  $f$ ,  $f : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}$ , with parameter  $\rho$ ,  $\rho \in \mathbb{R}_{++}$ , is defined as (Parikh and Boyd, 2014)

$$\text{prox}_{\rho^{-1}f}(\mathbf{V}) \triangleq \arg \min_{\mathbf{U} \in \mathbb{R}^{p \times p}} f(\mathbf{U}) + \frac{\rho}{2} \|\mathbf{U} - \mathbf{V}\|_{\text{F}}^2. \quad (69)$$

## Appendix C. Proofs

### C.1 Proof of Lemma 3

**Proof** We define an index set  $\Omega_t$ :

$$\Omega_t := \left\{ l \mid [\mathcal{L}\mathbf{w}]_{tt} = \sum_{l \in \Omega_t} x_l \right\}, \quad t \in [1, p]. \quad (70)$$

Then we have

$$\begin{aligned} \lambda_{\max}(\mathcal{L}^* \mathcal{L}) &= \sup_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathcal{L}^* \mathcal{L} \mathbf{x} = \sup_{\|\mathbf{x}\|=1} \|\mathcal{L}\mathbf{x}\|_F^2 = \sup_{\|\mathbf{x}\|=1} 2 \sum_{k=1}^{p(p-1)/2} x_k^2 + \sum_{i=1}^p ([\mathcal{L}\mathbf{w}]_{ii})^2 \\ &= \sup_{\|\mathbf{x}\|=1} 4 \sum_{k=1}^{p(p-1)/2} x_k^2 + \sum_{t=1}^p \sum_{i,j \in \Omega_t, i \neq j} x_i x_j \leq 4 + \sup_{\|\mathbf{x}\|=1} \frac{1}{2} \sum_{t=1}^p \sum_{i,j \in \Omega_t, i \neq j} x_i^2 + x_j^2 \\ &= (4 + 2(|\Omega_t| - 1)) \sup_{\|\mathbf{x}\|=1} \sum_{k=1}^{p(p-1)/2} x_k^2 = 2p, \end{aligned}$$

with equality if and only if  $x_1 = \dots = x_{p(p-1)/2} = \sqrt{\frac{2}{p(p-1)/2}}$  or  $x_1 = \dots = x_{p(p-1)/2} = -\sqrt{\frac{2}{p(p-1)/2}}$ . The last equality follows the fact that  $|\Omega_t| = p - 1$ .

Similarly, we can obtain

$$\begin{aligned} \lambda_{\max}(\mathbf{d}^* \mathbf{d}) &= \sup_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathbf{d}^* \mathbf{d} \mathbf{x} = \sup_{\|\mathbf{x}\|=1} \|\mathbf{d}\mathbf{x}\|^2 = \sup_{\|\mathbf{x}\|=1} 2 \sum_{k=1}^{p(p-1)/2} x_k^2 + \sum_{t=1}^p \sum_{i,j \in \Omega_t, i \neq j} x_i x_j \\ &\leq 2 + \sup_{\|\mathbf{x}\|=1} \frac{1}{2} \sum_{t=1}^p \sum_{i,j \in \Omega_t, i \neq j} x_i^2 + x_j^2 = (2 + 2(|\Omega_t| - 1)) \sup_{\|\mathbf{x}\|=1} \sum_{k=1}^{p(p-1)/2} x_k^2 = 2p - 2, \end{aligned}$$

with equality if and only if  $x_1 = \dots = x_{p(p-1)/2} = \sqrt{\frac{2}{p(p-1)/2}}$  or  $x_1 = \dots = x_{p(p-1)/2} = -\sqrt{\frac{2}{p(p-1)/2}}$ .

Finally, we have

$$\begin{aligned} \lambda_{\max}(\mathcal{L}^* \mathcal{L} + \mathbf{d}^* \mathbf{d}) &= \sup_{\|\mathbf{x}\|=1} \mathbf{x}^\top (\mathcal{L}^* \mathcal{L} + \mathbf{d}^* \mathbf{d}) \mathbf{x} \\ &\leq \sup_{\|\mathbf{x}\|=1} \mathbf{x}^\top (\mathcal{L}^* \mathcal{L}) \mathbf{x} + \sup_{\|\mathbf{y}\|=1} \mathbf{y}^\top (\mathbf{d}^* \mathbf{d}) \mathbf{y} \\ &= 4p - 2. \end{aligned} \quad (71)$$

Note that the equality in (71) can be achieved because the eigenvectors of  $\mathcal{L}^* \mathcal{L}$  and  $\mathbf{d}^* \mathbf{d}$  associated with the maximum eigenvalue are the same. Therefore, we conclude that  $\lambda_{\max}(\mathcal{L}^* \mathcal{L} + \mathbf{d}^* \mathbf{d}) = 4p - 2$ , completing the proof.  $\blacksquare$

## C.2 Proof of Theorem 4

We can rewrite the updating of  $\Theta$ ,  $\mathbf{w}$ ,  $\mathbf{Y}$  and  $\mathbf{y}$  in a compact form:

$$\Theta^{l+1} = \arg \min_{\Theta \succeq 0} L_\rho(\Theta, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l), \quad (72)$$

$$\mathbf{w}^{l+1} = \arg \min_{\mathbf{w} \geq 0} L_\rho(\Theta^{l+1}, \mathbf{w}, \mathbf{Y}^l, \mathbf{y}^l), \quad (73)$$

$$\begin{pmatrix} \mathbf{Y}^{l+1} \\ \mathbf{y}^{l+1} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}^l \\ \mathbf{y}^l \end{pmatrix} + \rho \begin{pmatrix} \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1} \\ \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \end{pmatrix}. \quad (74)$$

Now we can see that our ADMM algorithm satisfies the standard form with two blocks of primal variables  $\Theta$  and  $\mathbf{w}$ , and one block of dual variable  $(\mathbf{Y}, \mathbf{y})$ . Our ADMM approach splits the original problem into two blocks in (22). According to the existing convergence results of ADMM in (Boyd et al., 2011), we can conclude that Algorithm 2 will converge to the optimal primal-dual solution for (22).

## C.3 Proof of Theorem 5

**Proof** To prove Theorem 5, we first establish the boundedness of the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 3 in Lemma 16, and the monotonicity of  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)$  in Lemma 17.

**Lemma 16** *The sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 3 is bounded.*

**Proof** Let  $\mathbf{w}^0, \mathbf{V}^0, \mathbf{Y}^0$  and  $\mathbf{y}^0$  be the initialization of the sequences  $\{\mathbf{w}^l\}$ ,  $\{\mathbf{V}^l\}$ ,  $\{\mathbf{Y}^l\}$  and  $\{\mathbf{y}^l\}$ , respectively, and  $\|\mathbf{w}^0\|, \|\mathbf{V}^0\|_F, \|\mathbf{Y}^0\|_F$  and  $\|\mathbf{y}^0\|$  are bounded.

We prove the lemma by induction. Recall that the sequence  $\{\Theta^l\}$  is established by

$$\Theta^l = \frac{1}{2\rho} \mathbf{U}^{l-1} \left( \mathbf{\Gamma}^{l-1} + \sqrt{(\mathbf{\Gamma}^{l-1})^2 + 4\rho\mathbf{I}} \right) \mathbf{U}^{l-1\top}, \quad (75)$$

where  $\mathbf{\Gamma}^{l-1}$  contains the largest  $p - k$  eigenvalues of  $\rho\mathcal{L}\mathbf{w}^{l-1} - \mathbf{Y}^{l-1}$ , and  $\mathbf{U}^{l-1}$  contains the corresponding eigenvectors. When  $l = 1$ ,  $\|\mathbf{\Gamma}^0\|_F$  is bounded since both  $\|\mathbf{w}^0\|$  and  $\|\mathbf{Y}^0\|_F$  are bounded. Therefore, we can conclude that  $\|\Theta^1\|_F$  is bounded. The sequence  $\{\mathbf{w}^l\}$  is established by solving the subproblems

$$\begin{aligned} \mathbf{w}^l = \arg \min_{\mathbf{w} \geq 0} & \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} \\ & + \left\langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l \mathbf{V}^{l\top} - \mathbf{Y}^{l-1} - \rho \Theta^l \right) + \mathbf{d}^* \left( \mathbf{y}^{l-1} - \rho \mathbf{d} \right) \right\rangle. \end{aligned} \quad (76)$$

Let

$$f_l(\mathbf{w}) = \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} + \left\langle \mathbf{w}, \mathbf{a}^l \right\rangle,$$

where  $\mathbf{a}^l = \mathcal{L}^* (\mathbf{S} - \mathbf{Y}^{l-1} - \rho \Theta^l) + \mathbf{d}^* (\mathbf{y}^{l-1} - \rho \mathbf{d})$ . We get that  $\|\mathbf{a}^l\|$  is bounded because of the boundedness of  $\|\mathbf{Y}^0\|_F$ ,  $\|\mathbf{y}^0\|$ , and  $\|\Theta^1\|_F$ . By (Ying et al., 2020a), we know that  $\mathcal{L}^* \mathcal{L}$  is a positive definite matrix and the minimum eigenvalue  $\lambda_{\min}(\mathcal{L}^* \mathcal{L}) = 2$ . On the other hand,  $\mathbf{d}^* \mathbf{d}$  is a positive semi-definite matrix as follows,

$$\lambda_{\min}(\mathbf{d}^* \mathbf{d}) = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{d}^* \mathbf{d} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\langle \mathbf{d} \mathbf{x}, \mathbf{d} \mathbf{x} \rangle}{\mathbf{x}^\top \mathbf{x}} \geq 0. \quad (77)$$

Therefore, we obtain that

$$\lim_{\|\mathbf{w}\| \rightarrow +\infty} f_1(\mathbf{w}) \geq \lim_{\|\mathbf{w}\| \rightarrow +\infty} \rho \mathbf{w}^\top \mathbf{w} + \langle \mathbf{w}, \mathbf{a}^1 \rangle = +\infty,$$

implying that  $f_1(\mathbf{w})$  is coercive, *i.e.*,  $f_1(\mathbf{w}) \rightarrow +\infty$  for any  $\|\mathbf{w}\| \rightarrow +\infty$ . Thus  $\|\mathbf{w}^1\|$  is bounded. According to (48), we can see that  $\mathbf{V}$  is in a compact set, and thus  $\|\mathbf{V}^l\|_{\text{F}}$  is bounded for any  $l \geq 1$ . Finally, according to (36) and (37), one has

$$\mathbf{Y}^1 = \mathbf{Y}^0 + \rho (\Theta^1 - \mathcal{L} \mathbf{w}^1), \quad (78)$$

and

$$\mathbf{y}^1 = \mathbf{y}^0 + \rho (\mathbf{d} \mathbf{w}^1 - \mathbf{d}). \quad (79)$$

It is obvious that both  $\|\mathbf{Y}^1\|_{\text{F}}$  and  $\|\mathbf{y}^1\|$  are bounded. Therefore, it holds for  $l = 1$  that  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  is bounded.

Assume that  $\{(\Theta^{l-1}, \mathbf{w}^{l-1}, \mathbf{V}^{l-1}, \mathbf{Y}^{l-1}, \mathbf{y}^{l-1})\}$  is bounded for some  $l \geq 1$ , *i.e.*, each term in  $\{(\Theta^{l-1}, \mathbf{w}^{l-1}, \mathbf{V}^{l-1}, \mathbf{Y}^{l-1}, \mathbf{y}^{l-1})\}$  is bounded under  $\ell_2$ -norm or Frobenius norm. Following from (75), we can obtain that  $\|\Theta^l\|_{\text{F}}$  is bounded. By (76), similarly, we can get that  $\|\mathbf{w}^l\|$  is bounded. Similar to (78) and (79), we can also obtain that  $\|\mathbf{Y}^l\|$  and  $\|\mathbf{y}^l\|$  are bounded, because of the boundedness of  $\|\Theta^l\|_{\text{F}}$ ,  $\|\mathbf{w}^l\|$ ,  $\|\mathbf{Y}^{l-1}\|$  and  $\|\mathbf{y}^{l-1}\|$ . Thus,  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  is bounded, completing the induction. Therefore, we can conclude that the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  is bounded.  $\blacksquare$

**Lemma 17** *The sequence  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)$  generated by Algorithm 3 is lower bounded, and*

$$L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{V}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \leq L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l), \quad \forall l \in \mathbb{N}_+, \quad (80)$$

holds for any sufficiently large  $\rho$ .

**Proof** According to (43), we have

$$\begin{aligned} L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) &= \text{tr} \left( \mathcal{L} \mathbf{w}^l \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top \right) \right) - \log \det^* (\Theta^l) + \langle \mathbf{y}^l, \mathbf{d} \mathbf{w}^l - \mathbf{d} \rangle \\ &\quad + \frac{\rho}{2} \left\| \mathbf{d} \mathbf{w}^l - \mathbf{d} \right\|_2^2 + \langle \mathbf{Y}^l, \Theta^l - \mathcal{L} \mathbf{w}^l \rangle + \frac{\rho}{2} \left\| \Theta^l - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2. \end{aligned} \quad (81)$$

We can see that the lower boundedness of the sequence  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)$  can be established by the boundedness of  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  in Lemma 16.

Now we first establish that

$$L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) \leq L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l), \quad \forall l \in \mathbb{N}_+. \quad (82)$$

We have

$$\begin{aligned} L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) &= \text{tr} \left( \mathcal{L} \mathbf{w}^l \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top \right) \right) - \log \det^* (\Theta^{l+1}) + \langle \mathbf{y}^l, \mathbf{d} \mathbf{w}^l - \mathbf{d} \rangle \\ &\quad + \frac{\rho}{2} \left\| \mathbf{d} \mathbf{w}^l - \mathbf{d} \right\|_2^2 + \langle \mathbf{Y}^l, \Theta^{l+1} - \mathcal{L} \mathbf{w}^l \rangle + \frac{\rho}{2} \left\| \Theta^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2. \end{aligned}$$

Then we obtain

$$\begin{aligned} L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) &= -\log \det^*(\Theta^{l+1}) + \langle \mathbf{Y}^l, \Theta^{l+1} \rangle \\ &\quad + \frac{\rho}{2} \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^l\|_{\text{F}}^2 - \left( -\log \det^*(\Theta^l) + \langle \mathbf{Y}^l, \Theta^l \rangle + \frac{\rho}{2} \|\Theta^l - \mathcal{L}\mathbf{w}^l\|_{\text{F}}^2 \right). \end{aligned}$$

Note that  $\Theta^{l+1}$  minimizes the objective function

$$\Theta^{l+1} = \underset{\substack{\text{rank}(\Theta) = p-k \\ \Theta \succeq 0}}{\arg \min} -\log \det^*(\Theta) + \langle \Theta, \mathbf{Y}^l \rangle + \frac{\rho}{2} \|\Theta - \mathcal{L}\mathbf{w}^l\|_{\text{F}}^2. \quad (83)$$

Therefore

$$L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) \leq 0 \quad (84)$$

holds for any  $l \in \mathbb{N}_+$ .

One has

$$\begin{aligned} &L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{V}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \\ &= \underbrace{\text{tr} \left( \eta \mathcal{L}\mathbf{w}^l \left( \mathbf{V}^l \left( \mathbf{V}^l \right)^\top \right) \right) - \text{tr} \left( \eta \mathcal{L}\mathbf{w}^{l+1} \left( \mathbf{V}^{l+1} \left( \mathbf{V}^{l+1} \right)^\top \right) \right)}_{I_1} + \langle \mathcal{L}^* \mathbf{S}, \mathbf{w}^l - \mathbf{w}^{l+1} \rangle \\ &\quad + \underbrace{\langle \mathbf{y}^l, \mathbf{d}\mathbf{w}^l - \mathbf{d} \rangle - \langle \mathbf{y}^{l+1}, \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \rangle}_{I_2} + \underbrace{\langle \mathbf{Y}^l, \Theta^{l+1} - \mathcal{L}\mathbf{w}^l \rangle - \langle \mathbf{Y}^{l+1}, \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1} \rangle}_{I_3} \\ &\quad + \frac{\rho}{2} \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^l\|_{\text{F}}^2 - \frac{\rho}{2} \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}\|_{\text{F}}^2 + \frac{\rho}{2} \|\mathbf{d}\mathbf{w}^l - \mathbf{d}\|_2^2 - \frac{\rho}{2} \|\mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\|_2^2. \quad (85) \end{aligned}$$

The term  $I_1$  can be written as

$$\begin{aligned} I_1 &= \text{tr} \left( \eta \mathcal{L}\mathbf{w}^l \left( \mathbf{V}^l \left( \mathbf{V}^l \right)^\top \right) \right) - \text{tr} \left( \eta \mathcal{L}\mathbf{w}^{l+1} \left( \mathbf{V}^l \left( \mathbf{V}^l \right)^\top \right) \right) \\ &\quad + \underbrace{\text{tr} \left( \eta \mathcal{L}\mathbf{w}^{l+1} \left( \mathbf{V}^l \left( \mathbf{V}^l \right)^\top \right) \right) - \text{tr} \left( \eta \mathcal{L}\mathbf{w}^{l+1} \left( \mathbf{V}^{l+1} \left( \mathbf{V}^{l+1} \right)^\top \right) \right)}_{I_{1a}}. \end{aligned}$$

Note that  $\mathbf{V}^{l+1}$  is the optimal solution of the problem

$$\min_{\mathbf{V} \in \mathbb{R}^{p \times k}} \text{tr} \left( \mathbf{V}^\top \mathcal{L}\mathbf{w}^{l+1} \mathbf{V} \right), \quad \text{subject to } \mathbf{V}^\top \mathbf{V} = \mathbf{I}. \quad (86)$$

Thus the term  $I_{1a} \geq 0$ , and we can obtain

$$I_1 \geq \text{tr} \left( \eta \mathcal{L}\mathbf{w}^l \left( \mathbf{V}^l \left( \mathbf{V}^l \right)^\top \right) \right) - \text{tr} \left( \eta \mathcal{L}\mathbf{w}^{l+1} \left( \mathbf{V}^l \left( \mathbf{V}^l \right)^\top \right) \right). \quad (87)$$

For the term  $I_2$ , we have

$$\begin{aligned} I_2 &= \langle \mathbf{y}^l, \mathbf{d}\mathbf{w}^l - \mathbf{d} \rangle - \langle \mathbf{y}^l, \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \rangle - \rho \langle \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}, \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \rangle \\ &= \langle \mathbf{d}^* \mathbf{y}^l, \mathbf{w}^l - \mathbf{w}^{l+1} \rangle - \rho \|\mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\|_2^2, \end{aligned} \quad (88)$$

where the first equality is due to the updating of  $\mathbf{y}^{l+1}$  as below

$$\mathbf{y}^{l+1} = \mathbf{y}^l + \rho (\mathbf{d} \mathbf{w}^{l+1} - \mathbf{d}). \quad (89)$$

For the term  $I_3$ , similarly, we have

$$\begin{aligned} I_3 &= \langle \mathbf{Y}^l, \Theta^{l+1} - \mathcal{L}\mathbf{w}^l \rangle - \langle \mathbf{Y}^l, \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1} \rangle - \rho \langle \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}, \Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1} \rangle \\ &= \langle \mathcal{L}^* \mathbf{Y}^l, \mathbf{w}^{l+1} - \mathbf{w}^l \rangle - \rho \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}\|_F^2, \end{aligned} \quad (90)$$

where the first equality follows from

$$\mathbf{Y}^{l+1} = \mathbf{Y}^l + \rho (\Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}). \quad (91)$$

Therefore, we can obtain

$$I_2 + I_3 = \langle \mathbf{d}^* \mathbf{y}^l - \mathcal{L}^* \mathbf{Y}^l, \mathbf{w}^l - \mathbf{w}^{l+1} \rangle - \rho \|\mathbf{d} \mathbf{w}^{l+1} - \mathbf{d}\|_2^2 - \rho \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}\|_F^2. \quad (92)$$

Recall that  $\mathbf{w}^{l+1}$  is the optimal solution of the problem

$$\begin{aligned} \mathbf{w}^{l+1} &= \arg \min_{\mathbf{w} \geq \mathbf{0}} \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} \\ &\quad + \left\langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top - \mathbf{Y}^l - \rho \Theta^l \right) + \mathbf{d}^* (\mathbf{y}^l - \rho \mathbf{d}) \right\rangle. \end{aligned} \quad (93)$$

Thus,  $\mathbf{w}^{l+1}$  satisfies the KKT system of (93) as below

$$\rho (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w}^{l+1} + \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top - \mathbf{Y}^l - \rho \Theta^l \right) + \mathbf{d}^* (\mathbf{y}^l - \rho \mathbf{d}) - \boldsymbol{\nu} = \mathbf{0}; \quad (94)$$

$$w_i^{l+1} \nu_i = 0, \quad \text{for } i = 1, \dots, p(p-1)/2; \quad (95)$$

$$\mathbf{w}^{l+1} \geq \mathbf{0}, \quad \boldsymbol{\nu} \geq \mathbf{0}; \quad (96)$$

According to (94), we have

$$\mathbf{d}^* \mathbf{y}^l - \mathcal{L}^* \mathbf{Y}^l = -\rho (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w}^{l+1} - \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top - \rho \Theta^{l+1} \right) + \rho \mathbf{d}^* \mathbf{d} + \boldsymbol{\nu}. \quad (97)$$

Together with (92) and (97), we obtain

$$\begin{aligned} I_2 + I_3 &= \left\langle -\rho (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w}^{l+1} - \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top - \rho \Theta^{l+1} \right) + \rho \mathbf{d}^* \mathbf{d}, \mathbf{w}^l - \mathbf{w}^{l+1} \right\rangle \\ &\quad - \rho \|\mathbf{d} \mathbf{w}^{l+1} - \mathbf{d}\|_2^2 - \rho \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}\|_F^2 + \langle \boldsymbol{\nu}, \mathbf{w}^l - \mathbf{w}^{l+1} \rangle \\ &\geq \left\langle -\rho (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w}^{l+1} - \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^l (\mathbf{V}^l)^\top - \rho \Theta^{l+1} \right) + \rho \mathbf{d}^* \mathbf{d}, \mathbf{w}^l - \mathbf{w}^{l+1} \right\rangle \\ &\quad - \rho \|\mathbf{d} \mathbf{w}^{l+1} - \mathbf{d}\|_2^2 - \rho \|\Theta^{l+1} - \mathcal{L}\mathbf{w}^{l+1}\|_F^2, \end{aligned} \quad (98)$$

where the inequality is established by  $\langle \boldsymbol{\nu}, \mathbf{w}^l - \mathbf{w}^{l+1} \rangle \geq 0$ , which follows from the fact that  $\langle \boldsymbol{\nu}, \mathbf{w}^{l+1} \rangle = 0$  according to (95), and  $\langle \boldsymbol{\nu}, \mathbf{w}^l \rangle \geq 0$  because  $\boldsymbol{\nu} \geq \mathbf{0}$  by (96) and  $\mathbf{w}^l \geq \mathbf{0}$ . Plugging (87) and (98) into (85), by calculation we can obtain

$$\begin{aligned} & L_\rho(\boldsymbol{\Theta}^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\boldsymbol{\Theta}^{l+1}, \mathbf{w}^{l+1}, \mathbf{V}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \\ & \geq \frac{\rho}{2} \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\mathbf{w}^l \right\|_2^2 + \frac{\rho}{2} \left\| \mathcal{L}\mathbf{w}^{l+1} - \mathcal{L}\mathbf{w}^l \right\|_2^2 - \rho \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \right\|_2^2 - \rho \left\| \mathcal{L}\mathbf{w}^{l+1} - \boldsymbol{\Theta}^{l+1} \right\|_{\text{F}}^2 \\ & = \frac{\rho}{2} \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\mathbf{w}^l \right\|_2^2 + \frac{\rho}{2} \left\| \mathcal{L}\mathbf{w}^{l+1} - \mathcal{L}\mathbf{w}^l \right\|_{\text{F}}^2 - \frac{1}{\rho} \left\| \mathbf{y}^{l+1} - \mathbf{y}^l \right\|_2^2 - \frac{1}{\rho} \left\| \mathbf{Y}^{l+1} - \mathbf{Y}^l \right\|_{\text{F}}^2, \end{aligned} \quad (99)$$

where the equality follows from (89) and (91). If  $\rho$  is sufficiently large such that

$$\rho \geq \max_l \left( \frac{2c \left( \left\| \mathbf{y}^{l+1} - \mathbf{y}^l \right\|_2^2 + \left\| \mathbf{Y}^{l+1} - \mathbf{Y}^l \right\|_{\text{F}}^2 \right)}{\left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\mathbf{w}^l \right\|_2^2 + \left\| \mathcal{L}\mathbf{w}^{l+1} - \mathcal{L}\mathbf{w}^l \right\|_{\text{F}}^2} \right)^{\frac{1}{2}} \quad (100)$$

holds with some constant  $c > 1$ , together with (84), we can conclude that

$$L_\rho(\boldsymbol{\Theta}^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) \geq L_\rho(\boldsymbol{\Theta}^{l+1}, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) \geq L_\rho(\boldsymbol{\Theta}^{l+1}, \mathbf{w}^{l+1}, \mathbf{V}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}),$$

for any  $l \in \mathbb{N}_+$ . Note that besides a fixed parameter  $\rho$ , an alternative strategy is to increase the  $\rho$  iteratively (Ying et al., 2017) until the condition (100) could be satisfied well. ■

Now we are ready to prove Theorem 5. By Lemma 16, the sequence  $\{(\boldsymbol{\Theta}^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  is bounded. Therefore, there exists at least one convergent subsequence  $\{(\boldsymbol{\Theta}^{l_s}, \mathbf{w}^{l_s}, \mathbf{V}^{l_s}, \mathbf{Y}^{l_s}, \mathbf{y}^{l_s})\}_{s \in \mathbb{N}}$ , which converges to a limit point denoted by  $\{(\boldsymbol{\Theta}^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{V}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty})\}$ . By Lemma 17, we obtain that  $L_\rho(\boldsymbol{\Theta}^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)$  is monotonically decreasing and lower bounded, and thus is convergent. Note that the function  $\log \det^*(\boldsymbol{\Theta})$  is continuous over the set  $\mathcal{S} = \{\boldsymbol{\Theta} \in \mathcal{S}_+^p \mid \text{rank}(\boldsymbol{\Theta}) = p - k\}$ . We can get

$$\lim_{l \rightarrow +\infty} L_\rho(\boldsymbol{\Theta}^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) = L_\rho(\boldsymbol{\Theta}^\infty, \mathbf{w}^\infty, \mathbf{V}^\infty, \mathbf{Y}^\infty, \mathbf{y}^\infty) = L_\rho(\boldsymbol{\Theta}^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{V}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty}).$$

The (99), (100) and (101) together yields

$$\begin{aligned} & L_\rho(\boldsymbol{\Theta}^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\boldsymbol{\Theta}^{l+1}, \mathbf{w}^{l+1}, \mathbf{V}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \\ & \geq (c - 1)\rho \left( \left\| \mathcal{L}\mathbf{w}^{l+1} - \boldsymbol{\Theta}^{l+1} \right\|_{\text{F}}^2 + \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \right\|_2^2 \right). \end{aligned} \quad (101)$$

Thus, we obtain

$$\lim_{l \rightarrow +\infty} \left\| \mathcal{L}\mathbf{w}^l - \boldsymbol{\Theta}^l \right\|_{\text{F}} = 0, \quad \text{and} \quad \lim_{l \rightarrow +\infty} \left\| \mathbf{d}\mathbf{w}^l - \mathbf{d} \right\|_2 = 0. \quad (102)$$

Obviously,  $\left\| \mathcal{L}\mathbf{w}^{l_s} - \boldsymbol{\Theta}^{l_s} \right\|_{\text{F}} \rightarrow 0$  and  $\left\| \mathbf{d}\mathbf{w}^{l_s} - \mathbf{d} \right\|_2 \rightarrow 0$  also hold for any subsequence as  $s \rightarrow +\infty$ , which implies that  $\mathbf{Y}^{l_\infty}$  and  $\mathbf{y}^{l_\infty}$  satisfy the condition of stationary point of  $L_\rho(\boldsymbol{\Theta}, \mathbf{w}, \mathbf{V}, \mathbf{Y}, \mathbf{y})$  with respect to  $\mathbf{Y}$  and  $\mathbf{y}$ , respectively. By (91) and (89), we also have

$$\lim_{l \rightarrow +\infty} \left\| \mathbf{Y}^{l+1} - \mathbf{Y}^l \right\|_{\text{F}} = 0, \quad \text{and} \quad \lim_{l \rightarrow +\infty} \left\| \mathbf{y}^{l+1} - \mathbf{y}^l \right\|_2 = 0. \quad (103)$$

Together with (99), we obtain

$$\lim_{l \rightarrow +\infty} \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\mathbf{w}^l \right\|_2 = 0 \quad \text{and} \quad \left\| \mathcal{L}\mathbf{w}^{l+1} - \mathcal{L}\mathbf{w}^l \right\|_{\text{F}} = 0. \quad (104)$$

Recall that  $\mathbf{V}^l$  contains the  $k$  eigenvectors associated with the  $k$  smallest eigenvalues of  $\mathcal{L}\mathbf{w}^l$ , and thus it is easy to check that

$$\lim_{l \rightarrow +\infty} \left\| \mathbf{V}^{l+1} - \mathbf{V}^l \right\|_{\text{F}} = 0. \quad (105)$$

For the limit point  $\{(\Theta^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{V}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty})\}$  of any subsequence  $\{(\Theta^{l_s}, \mathbf{w}^{l_s}, \mathbf{V}^{l_s}, \mathbf{Y}^{l_s}, \mathbf{y}^{l_s})\}_{s \in \mathbb{N}}$ ,  $\Theta^{l_\infty}$  minimizes the following subproblem

$$\begin{aligned} \Theta^{l_\infty} &= \arg \min_{\substack{\text{rank}(\Theta) = p-k \\ \Theta \succeq 0}} -\log \det^*(\Theta) + \langle \Theta, \mathbf{Y}^{l_\infty-1} \rangle + \frac{\rho}{2} \left\| \Theta - \mathcal{L}\mathbf{w}^{l_\infty-1} \right\|_{\text{F}}^2 \\ &= \arg \min_{\substack{\text{rank}(\Theta) = p-k \\ \Theta \succeq 0}} -\log \det^*(\Theta) + \langle \Theta, \mathbf{Y}^{l_\infty} \rangle + \frac{\rho}{2} \left\| \Theta - \mathcal{L}\mathbf{w}^{l_\infty} \right\|_{\text{F}}^2 \\ &\quad - \langle \Theta, \mathbf{Y}^{l_\infty} - \mathbf{Y}^{l_\infty-1} \rangle + \rho \langle \Theta, \mathcal{L}\mathbf{w}^{l_\infty} - \mathcal{L}\mathbf{w}^{l_\infty-1} \rangle. \end{aligned}$$

By (103) and (104), we conclude that  $\Theta^{l_\infty}$  satisfies the condition of stationary point of  $L_\rho(\Theta, \mathbf{w}, \mathbf{V}, \mathbf{Y}, \mathbf{y})$  with respect to  $\Theta$ . Similarly,  $\mathbf{w}^{l_\infty}$  minimizes the subproblem

$$\begin{aligned} \mathbf{w}^{l_\infty} &= \arg \min_{\mathbf{w} \geq 0} \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} + \langle \mathbf{w}, \mathbf{d}^* (\mathbf{y}^{l_\infty-1} - \rho \mathbf{d}) \rangle \\ &\quad + \langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^{l_\infty-1} (\mathbf{V}^{l_\infty-1})^\top - \mathbf{Y}^{l_\infty-1} - \rho \Theta^{l_\infty} \right) \rangle \\ &= \arg \min_{\mathbf{w} \geq 0} \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} + \langle \mathbf{w}, \mathbf{d}^* (\mathbf{y}^{l_\infty} - \rho \mathbf{d}) \rangle \\ &\quad + \langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{S} + \eta \mathbf{V}^{l_\infty} (\mathbf{V}^{l_\infty})^\top - \mathbf{Y}^{l_\infty} - \rho \Theta^{l_\infty} \right) \rangle + \langle \mathcal{L}\mathbf{w}, \mathbf{Y}^{l_\infty} - \mathbf{Y}^{l_\infty-1} \rangle \\ &\quad + \langle \mathbf{d}\mathbf{w}, \mathbf{y}^{l_\infty-1} - \mathbf{y}^{l_\infty} \rangle + \eta \langle \mathcal{L}\mathbf{w}, \mathbf{V}^{l_\infty-1} (\mathbf{V}^{l_\infty-1})^\top - \mathbf{V}^{l_\infty} (\mathbf{V}^{l_\infty})^\top \rangle. \end{aligned}$$

By (103), (104) and (105),  $\mathbf{w}^{l_\infty}$  satisfies the condition of stationary point of  $L_\rho(\Theta, \mathbf{w}, \mathbf{V}, \mathbf{Y}, \mathbf{y})$  with respect to  $\mathbf{w}$ .  $\mathbf{V}^{l_\infty}$  minimizes the subproblem

$$\mathbf{V}^{l_\infty} = \arg \min_{\mathbf{V} \in \mathbb{R}^{p \times k}} \text{tr} \left( \mathbf{V}^\top \mathcal{L}\mathbf{w}^{l_\infty} \mathbf{V} \right), \quad \text{subject to } \mathbf{V}^\top \mathbf{V} = \mathbf{I},$$

which implies that  $\mathbf{V}^{l_\infty}$  satisfies the condition of stationary point of  $L_\rho(\Theta, \mathbf{w}, \mathbf{V}, \mathbf{Y}, \mathbf{y})$  with respect to  $\mathbf{V}$ . To sum up, we can conclude that any limit point  $\{(\Theta^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{V}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty})\}$  of the sequence generated by Algorithm 3 is a stationary point of  $L_\rho(\Theta, \mathbf{w}, \mathbf{V}, \mathbf{Y}, \mathbf{y})$ .  $\blacksquare$

#### C.4 Proof of Theorem 6

**Proof** Similar to the proof of Theorem 5, we establish the boundedness of the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 4 in Lemma 18, and the monotonicity of  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)$  in Lemma 19.

**Lemma 18** *The sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 4 is bounded.*

**Proof** Let  $\mathbf{w}^0, \mathbf{Y}^0$  and  $\mathbf{y}^0$  be the initialization of the sequences  $\{\mathbf{w}^l\}$ ,  $\{\mathbf{Y}^l\}$  and  $\{\mathbf{y}^l\}$ , respectively, and  $\|\mathbf{w}^0\|, \|\mathbf{Y}^0\|_F$  and  $\|\mathbf{y}^0\|$  are bounded.

We prove the boundedness of the sequence by induction. Notice that the subproblem for  $\Theta$  is the same with that in (75), and thus we can directly get the boundedness of  $\|\Theta^l\|_F$  for  $l = 1$ . The sequence  $\{\mathbf{w}^l\}$  is established by solving the subproblems

$$\begin{aligned} \min_{\mathbf{w} \geq \mathbf{0}} \quad & \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} - \left\langle \mathbf{w}, \mathcal{L}^* \left( \mathbf{Y}^{l-1} + \rho \Theta^l \right) - \mathbf{d}^* \left( \mathbf{y}^{l-1} - \rho \mathbf{d} \right) \right\rangle \\ & + \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right). \end{aligned} \quad (106)$$

Let

$$g_l(\mathbf{w}) = \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w} + \langle \mathbf{w}, \mathbf{a}^l \rangle + \frac{p + \nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right), \quad (107)$$

where  $\mathbf{a}^l = \mathcal{L}^* (\mathbf{S} - \mathbf{Y}^{l-1} - \rho \Theta^l) + \mathbf{d}^* (\mathbf{y}^{l-1} - \rho \mathbf{d})$ . Note that  $\|\mathbf{a}^l\|$  is bounded because  $\|\mathbf{Y}^0\|_F, \|\mathbf{y}^0\|$ , and  $\|\Theta^1\|_F$  are bounded. In the proof of Lemma 16, we have shown that  $\mathcal{L}^* \mathcal{L}$  is a positive definite matrix with the minimum eigenvalue  $\lambda_{\min}(\mathcal{L}^* \mathcal{L}) = 2$ , and  $\mathbf{d}^* \mathbf{d}$  is a positive semi-definite matrix. Since  $\log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w} \mathbf{x}_{i,*}}{\nu} \right) \geq 0$  for any  $\mathbf{w} \geq \mathbf{0}$ , we have

$$\lim_{\|\mathbf{w}\| \rightarrow +\infty} g_1(\mathbf{w}) \geq \lim_{\|\mathbf{w}\| \rightarrow +\infty} \rho \mathbf{w}^\top \mathbf{w} + \langle \mathbf{w}, \mathbf{a}^1 \rangle = +\infty. \quad (108)$$

Thus  $g_1(\mathbf{w})$  is coercive. Recall that we solve the optimization (106) by the MM framework. Hence, the objective function value is monotonically decreasing as a function of the iterations, and  $\mathbf{w}^l$  is a stationary point of (106). Then the coercivity of  $g_1(\mathbf{w})$  yields the boundedness of  $\|\mathbf{w}^l\|$ . Finally,  $\|\mathbf{Y}^l\|_F$  and  $\|\mathbf{y}^l\|$  are also bounded, because  $\mathbf{Y}^l$  and  $\mathbf{y}^l$  are updated as done in (78) and (79), respectively, and thus the proof is the same.

Now we assume that  $\{(\Theta^{l-1}, \mathbf{w}^{l-1}, \mathbf{Y}^{l-1}, \mathbf{y}^{l-1})\}$  is bounded for some  $l \geq 1$ , and check the boundedness of  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$ . Similar to the proof in (75), we can prove that  $\|\Theta^l\|_F$  is bounded. By (106), we can also obtain the boundedness of  $\|\mathbf{w}^l\|$ . We can also obtain that  $\|\mathbf{Y}^l\|$  and  $\|\mathbf{y}^l\|$  are bounded according to the boundedness of  $\|\Theta^l\|_F, \|\mathbf{w}^l\|, \|\mathbf{Y}^{l-1}\|$  and  $\|\mathbf{y}^{l-1}\|$ . Thus,  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  is bounded, completing the induction. Therefore, we establish the boundedness of the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$ .  $\blacksquare$

**Lemma 19** *The sequence  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)$  generated by Algorithm 4 is lower bounded, and*

$$L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \leq L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l), \quad \forall l \in \mathbb{N}_+, \quad (109)$$

holds for any sufficiently large  $\rho$ .

**Proof** According to (50), we have

$$\begin{aligned} L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) &= \frac{p+\nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w}^l \mathbf{x}_{i,*}}{\nu} \right) - \log \det(\Theta^l + \mathbf{J}) + \langle \mathbf{y}^l, \mathbf{d} - \mathbf{w}^l \rangle \\ &\quad + \frac{\rho}{2} \left\| \mathbf{d} - \mathbf{w}^l \right\|_2^2 + \langle \mathbf{Y}^l, \Theta^l - \mathcal{L} \mathbf{w}^l \rangle + \frac{\rho}{2} \left\| \Theta^l - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2. \end{aligned} \quad (110)$$

We can see that the lower boundedness of the sequence  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)$  can be established by the boundedness of  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  in Lemma 18.

By a similar argument in the proof of Lemma 17, we can also establish that

$$L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) \leq L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l), \quad \forall l \in \mathbb{N}_+. \quad (111)$$

One has

$$\begin{aligned} &L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \\ &= \underbrace{\langle \mathbf{y}^l, \mathbf{d} - \mathbf{w}^l \rangle - \langle \mathbf{y}^{l+1}, \mathbf{d} - \mathbf{w}^{l+1} \rangle}_{I_1} + \underbrace{\langle \mathbf{Y}^l, \Theta^{l+1} - \mathcal{L} \mathbf{w}^l \rangle - \langle \mathbf{Y}^{l+1}, \Theta^{l+1} - \mathcal{L} \mathbf{w}^{l+1} \rangle}_{I_2} \\ &\quad + r(\mathcal{L} \mathbf{w}^l) - r(\mathcal{L} \mathbf{w}^{l+1}) + \frac{\rho}{2} \left\| \mathbf{d} - \mathbf{w}^l \right\|_2^2 - \frac{\rho}{2} \left\| \mathbf{d} - \mathbf{w}^{l+1} \right\|_2^2 \\ &\quad + \frac{\rho}{2} \left\| \Theta^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2 - \frac{\rho}{2} \left\| \Theta^{l+1} - \mathcal{L} \mathbf{w}^{l+1} \right\|_{\text{F}}^2, \end{aligned} \quad (112)$$

where  $r(\mathbf{L}) = \frac{p+\nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathbf{L} \mathbf{x}_{i,*}}{\nu} \right)$ . According to (88) and (90), we obtain

$$I_1 + I_2 = \langle \mathbf{d} - \mathbf{w}^l - \mathcal{L}^* \mathbf{Y}^l, \mathbf{w}^l - \mathbf{w}^{l+1} \rangle - \rho \left\| \mathbf{d} - \mathbf{w}^{l+1} \right\|_2^2 - \rho \left\| \Theta^{l+1} - \mathcal{L} \mathbf{w}^{l+1} \right\|_{\text{F}}^2. \quad (113)$$

According to the convergence result of the majorization-minimization framework (Sun et al., 2017), we know that any limit point of the sequence is a stationary point of the following problem

$$\min_{\mathbf{w} \geq \mathbf{0}} \frac{p+\nu}{n} \sum_{i=1}^n \log \left( 1 + \frac{\mathbf{x}_{i,*}^\top \mathcal{L} \mathbf{w}^l \mathbf{x}_{i,*}}{\nu} \right) + \frac{\rho}{2} \mathbf{w}^\top (\mathbf{d} - \mathcal{L}^* \mathcal{L}) \mathbf{w} - \langle \mathbf{w}, \mathcal{L}^* (\mathbf{Y}^l + \rho \Theta^l) - \mathbf{d} \rangle. \quad (114)$$

The set of the stationary points for the optimization (114) is defined by

$$\mathcal{X} = \left\{ \mathbf{w} \mid \nabla g_l(\mathbf{w})^\top (\mathbf{z} - \mathbf{w}) \geq 0, \forall \mathbf{z} \geq \mathbf{0} \right\}, \quad (115)$$

where  $g_l(\mathbf{w})$  is the objective function in (114). The existence of the limit point can be guaranteed by the coercivity of  $g_l(\mathbf{w})$ , which has been established in the proof of Lemma 18. Therefore,  $\mathbf{w}^{l+1}$  is a stationary point. By taking  $\mathbf{z} = \mathbf{w}^l$  and  $\mathbf{w} = \mathbf{w}^{l+1}$  in (115), we obtain

$$\left( \mathcal{L}^* \left( \nabla r \left( \mathcal{L} \mathbf{w}^{l+1} \right) \right) + \rho (\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w}^{l+1} - \mathcal{L}^* \left( \mathbf{Y}^l + \rho \Theta^l \right) + \mathbf{d}^* \left( \mathbf{y}^l - \rho \mathbf{d} \right) \right)^\top \left( \mathbf{w}^l - \mathbf{w}^{l+1} \right) \geq 0.$$

Thus, we have

$$\begin{aligned} \left\langle \mathbf{d}^* \mathbf{y}^l - \mathcal{L}^* \mathbf{Y}^l, \mathbf{w}^l - \mathbf{w}^{l+1} \right\rangle &\geq - \left\langle \nabla r \left( \mathcal{L} \mathbf{w}^{l+1} \right), \mathcal{L} \mathbf{w}^l - \mathcal{L} \mathbf{w}^{l+1} \right\rangle \\ &\quad + \rho \left\langle -(\mathbf{d}^* \mathbf{d} + \mathcal{L}^* \mathcal{L}) \mathbf{w}^{l+1} + \mathcal{L}^* \Theta^l + \mathbf{d}^* \mathbf{d}, \mathbf{w}^l - \mathbf{w}^{l+1} \right\rangle, \end{aligned} \quad (116)$$

Plugging (113) and (116) into (112), we obtain

$$\begin{aligned} &L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \\ &\geq \frac{\rho}{2} \left\| \mathbf{d} \mathbf{w}^{l+1} - \mathbf{d} \mathbf{w}^l \right\|_2^2 + \frac{\rho}{2} \left\| \mathcal{L} \mathbf{w}^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_2^2 - \rho \left\| \mathbf{d} \mathbf{w}^{l+1} - \mathbf{d} \right\|_2^2 - \rho \left\| \mathcal{L} \mathbf{w}^{l+1} - \Theta^{l+1} \right\|_{\text{F}}^2 \\ &\quad + r(\mathcal{L} \mathbf{w}^l) - r(\mathcal{L} \mathbf{w}^{l+1}) - \left\langle \nabla r \left( \mathcal{L} \mathbf{w}^{l+1} \right), \mathcal{L} \mathbf{w}^l - \mathcal{L} \mathbf{w}^{l+1} \right\rangle \\ &\geq \frac{\rho}{2} \left\| \mathbf{d} \mathbf{w}^{l+1} - \mathbf{d} \mathbf{w}^l \right\|_2^2 + \frac{\rho - L_r}{2} \left\| \mathcal{L} \mathbf{w}^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2 - \frac{1}{\rho} \left\| \mathbf{y}^{l+1} - \mathbf{y}^l \right\|_2^2 - \frac{1}{\rho} \left\| \mathbf{Y}^{l+1} - \mathbf{Y}^l \right\|_{\text{F}}^2, \end{aligned} \quad (117)$$

where the last inequality is due to the fact that  $r(\mathbf{L})$  is a concave function and has  $L_r$ -Lipschitz continuous gradient, in which  $L_r > 0$  is a constant, thus we have

$$r(\mathcal{L} \mathbf{w}^l) - r(\mathcal{L} \mathbf{w}^{l+1}) - \left\langle \nabla r \left( \mathcal{L} \mathbf{w}^{l+1} \right), \mathcal{L} \mathbf{w}^l - \mathcal{L} \mathbf{w}^{l+1} \right\rangle \geq -\frac{L_r}{2} \left\| \mathcal{L} \mathbf{w}^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2. \quad (118)$$

By calculation, we obtain that if  $\rho$  is sufficiently large such that

$$\rho \geq \max \left( L_r, \max_l \frac{L_r \left\| \mathcal{L} \mathbf{w}^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2 + \left( L_r^2 \left\| \mathcal{L} \mathbf{w}^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^4 + 8abc \right)^{\frac{1}{2}}}{2a} \right) \quad (119)$$

holds with some constant  $c > 1$ , where  $a = \left\| \mathbf{d} \mathbf{w}^{l+1} - \mathbf{d} \mathbf{w}^l \right\|_2^2 + \left\| \mathcal{L} \mathbf{w}^{l+1} - \mathcal{L} \mathbf{w}^l \right\|_{\text{F}}^2$  and  $b = \left\| \mathbf{y}^{l+1} - \mathbf{y}^l \right\|_2^2 + \left\| \mathbf{Y}^{l+1} - \mathbf{Y}^l \right\|_{\text{F}}^2$ , then, together with (111), we conclude that

$$L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) \geq L_\rho(\Theta^{l+1}, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) \geq L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}),$$

for any  $l \in \mathbb{N}_+$ . Note that for the case of Gaussian distribution in Section 4.4, the constant  $L_r$  will be zero, and  $\rho$  in (119) will be consistent with that in (100).  $\blacksquare$

Now we are ready to prove Theorem 6. By Lemma 18, the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 4 is bounded. Therefore, there exists at least one convergent subsequence  $\{(\Theta^{l_s}, \mathbf{w}^{l_s}, \mathbf{Y}^{l_s}, \mathbf{y}^{l_s})\}_{s \in \mathbb{N}}$ , which converges to the limit point denoted by  $\{(\Theta^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty})\}$ .

By Lemma 19, we obtain that the sequence  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)$  defined in (110) is monotonically decreasing and lower bounded, implying that  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)$  is convergent. We can get  $\lim_{l \rightarrow +\infty} L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) = L_\rho(\Theta^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty})$ . The (117), (119) and (120) together yields

$$\begin{aligned} L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l) - L_\rho(\Theta^{l+1}, \mathbf{w}^{l+1}, \mathbf{Y}^{l+1}, \mathbf{y}^{l+1}) \\ \geq (c-1)\rho \left( \left\| \mathcal{L}\mathbf{w}^{l+1} - \Theta^{l+1} \right\|_{\text{F}}^2 + \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d} \right\|_2^2 \right). \end{aligned} \quad (120)$$

The convergence of  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{Y}^l, \mathbf{y}^l)$  yields

$$\lim_{l \rightarrow +\infty} \left\| \mathcal{L}\mathbf{w}^l - \Theta^l \right\|_{\text{F}} = 0, \quad \text{and} \quad \lim_{l \rightarrow +\infty} \left\| \mathbf{d}\mathbf{w}^l - \mathbf{d} \right\|_2 = 0. \quad (121)$$

By the updating of  $\mathbf{Y}^{l+1}$  and  $\mathbf{y}^{l+1}$ , we can get

$$\lim_{l \rightarrow +\infty} \left\| \mathbf{Y}^{l+1} - \mathbf{Y}^l \right\|_{\text{F}} = 0, \quad \text{and} \quad \lim_{l \rightarrow +\infty} \left\| \mathbf{y}^{l+1} - \mathbf{y}^l \right\|_2 = 0. \quad (122)$$

Together with (117), we obtain

$$\lim_{l \rightarrow +\infty} \left\| \mathbf{d}\mathbf{w}^{l+1} - \mathbf{d}\mathbf{w}^l \right\|_2 = 0 \quad \text{and} \quad \left\| \mathcal{L}\mathbf{w}^{l+1} - \mathcal{L}\mathbf{w}^l \right\|_{\text{F}} = 0. \quad (123)$$

Similar to the proof of Theorem 5, by (121), (122) and (123), we can prove that  $\Theta^{l_\infty}$ ,  $\mathbf{w}^{l_\infty}$ ,  $\mathbf{Y}^{l_\infty}$  and  $\mathbf{y}^{l_\infty}$  satisfy the condition of stationary point of  $L_\rho(\Theta, \mathbf{w}, \mathbf{V}, \mathbf{Y}, \mathbf{y})$  with respect to  $\Theta$ ,  $\mathbf{w}$ ,  $\mathbf{Y}$  and  $\mathbf{y}$ , respectively. To sum up, we can conclude that any limit point  $\{(\Theta^{l_\infty}, \mathbf{w}^{l_\infty}, \mathbf{Y}^{l_\infty}, \mathbf{y}^{l_\infty})\}$  of the sequence is a stationary point of  $L_\rho(\Theta, \mathbf{w}, \mathbf{Y}, \mathbf{y})$ .  $\blacksquare$

## C.5 Proof of Theorem 7

**Proof** The proof of Theorem 7 is similar to the proof of Theorems 5 and 6. We can establish the boundedness of the sequence  $\{(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)\}$  generated by Algorithm 5 as done in Lemma 16 and 18. Similar to Lemmas 17 and 19, we can establish the monotonicity and boundedness of  $L_\rho(\Theta^l, \mathbf{w}^l, \mathbf{V}^l, \mathbf{Y}^l, \mathbf{y}^l)$ . Therefore, we omit the details of the proof of Theorem 7 to avoid redundancy.  $\blacksquare$

## References

- P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2007.
- R. Agrawal, U. Roy, and C. Uhler. Covariance Matrix Estimation under Total Positivity for Portfolio Selection. *Journal of Financial Econometrics*, 09 2020.
- A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky. High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion. *Journal of Machine Learning Research*, 13(1):2293–2337, 2012.
- O. Banerjee, L. E. Ghaoui, and A. d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(15):485–516, 2008.
- G. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E*, 68, 2003.
- G. Bonanno, G. Caldarelli, F. Lillo, S. Micciché, N. Vandewalle, and R. N. Mantegna. Networks of equities in financial markets. *The European Physical Journal B*, 38:363–371, 2004.
- S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- G. Carlsson and F. Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 2010.
- Z. Chen, L. Li, and J. Bruna. Supervised community detection with graph neural networks. In *International Conference on Learning Representations (ICLR’19)*, 2019.
- S. P. Chepuri, S. Liu, G. Leus, and A. O. Hero. Learning sparse graphs under smoothness prior. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6508–6512, 2017.
- F. R. K. Chung. *Spectral Graph Theory*, volume 92. CBMS Regional Conference Series in Mathematics, 1997.
- R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 2001.
- M. Coutino, E. Isufi, T. Maehara, and G. Leus. State-space network topology identification from partial observations. In *arXiv: 1906.10471*, 2019.
- G. Csárdi. *igraph*: Network analysis and visualization. *CRAN Vignette*, 2019.
- D. Eddelbuettel D. Bates. Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, 52, 2013.

- P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B*, 76(2):373–397, 2014.
- J. V. de M. Cardoso and D. P. Palomar. Learning undirected graphs in financial markets. In *54th Annual Asilomar Conference on Signals, Systems, and Computers*, 2020.
- M. L. de Prado. Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management*, 42(4):59–69, 2016.
- M. L. de Prado. *Machine Learning for Asset Managers (Elements in Quantitative Finance)*. Cambridge University Press, 2020.
- S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst. Learning Laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016.
- P. Donnat, G. Marti, and P. Very. Toward a generic representation of random variables for machine learning. *Pattern Recognition Letters*, 70:24–31, 2016.
- C. Dose and S. Cincotti. Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*, 355(1):145 – 151, 2005.
- D. Eddelbuettel and R. Francois. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40, 2011.
- D. Eddelbuettel and C. Sanderson. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, 71, 2014.
- H. E. Egilmez, E. Pavez, and A. Ortega. Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, 2017.
- E. F. Fama and K. R. French. *Journal of Economic Perspective*, 18(3):25–46, 2004.
- K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.
- Y. Feng and D. Palomar. A signal processing perspective on financial engineering. *Foundations and Trends in Signal Processing*, 9:1–231, 2015.
- S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010.
- J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–41, 2008.

- A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software, Articles*, 94(14):1–34, 2020. ISSN 1548-7660.
- C. Gourieroux and A. Monfort. *Time Series and Dynamic Models*. Themes in Modern Econometrics. Cambridge University Press, 1997.
- B. Hao, W. W. Sun, Y. Liu, and G. Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, 18(217):1–58, 2018.
- A. C. Harvey. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Cambridge University Press, 2013.
- S. Hassan-Moghaddam, N. K. Dhingra, and M. R. Jovanović. Topology identification of undirected consensus networks via sparse inverse covariance estimation. In *IEEE 55th Conference on Decision and Control (CDC)*, pages 4624–4629, 2016.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- C. Hsieh, A. Banerjee, I. S. Dhillon, and P. K. Ravikumar. A divide-and-conquer method for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems (NeurIPS'12)*, pages 2330–2338, 2012.
- V. Kalofolias. How to learn a graph from smooth signals. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 920–929, 2016.
- M. Kazakov and V. A. Kalyagin. Spectral properties of financial correlation matrices. In *Models, Algorithms and Technologies for Network Analysis*, pages 135–156, Cham, 2016.
- A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.
- O. Knill. Cauchy–Binet for pseudo-determinants. *Linear Algebra and its Applications*, 459: 522 – 547, 2014.
- N. Komodakis and J. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015.
- S. Kumar, J. Ying, J. V. de M. Cardoso, and D. P. Palomar. Structured graph learning via laplacian spectral constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.
- S. Kumar, J. Ying, J. V. de M. Cardoso, and D. P. Palomar. Bipartite structured Gaussian graphical modeling via adjacency spectral priors. In *53rd Annual Asilomar Conference on Signals, Systems, and Computers*, 2019b.
- S. Kumar, J. Ying, J. V. de M. Cardoso, and D. P. Palomar. A unified framework for structured graph learning via spectral constraints. *Journal of Machine Learning Research*, 21:1–60, 2020.

- B. M. Lake and J. B. Tenenbaum. Discovering structure by learning sparse graph. In *Proceedings of the 33rd Annual Cognitive Science Conference*, 2010.
- L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(3):391–397, 2000.
- V. Lemieux, P. S. Rahmdel, R. Walker, B. L. W. Wong, and M. Flood. Clustering techniques and their effect on portfolio formation and risk analysis. In *Proceedings of the International Workshop on Data Science for Macro-Modeling*, page 1–6, 2014.
- Y. Li, C. Sha, X. Huang, and Y. Zhang. Community detection in attributed graphs: An embedding approach. In *AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.
- J. Liu, S. Kumar, and D. P. Palomar. Parameter estimation of heavy-tailed ar model with missing data via stochastic em. *IEEE Transactions on Signal Processing*, 67(8):2159–2172, 2019.
- Y. Malevergne and D. Sornette. *Extreme Financial Risks: From Dependence to Risk Management*. Springer-Verlag, 2006.
- R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999.
- R. N. Mantegna and H. E. Stanley. *An Introduction to Econophysics: Correlation and Complexity in Finance*. Cambridge University Press, 2004.
- A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro. Sampling of graph signals with successive local aggregations. *IEEE Transactions on Signal Processing*, 64(7):1832–1843, 2016.
- G. Marti, P. Very, P. Donnat, and F. Nielsen. A proposal of a methodological framework with experimental guidelines to investigate clustering stability on financial time series. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 32–37, 2015.
- G. Marti, S. Andler, F. Nielsen, and P. Donnat. Clustering financial time series: How long is enough? In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016.
- G. Marti, F. Nielsen, M. Biríkowskí, and P. Donnat. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. In *arXiv: 1703.00485*, 2017a.
- G. Marti, F. Nielsen, P. Donnat, and S. Andler. On clustering financial time series: a need for distances between dependent random variables. *Computational Information Geometry*, pages 149–174, 2017b.
- G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3):16–43, 2019.

- T. Millington and M. Niranjan. Partial correlation financial networks. *Applied Network Science*, 5, 2020.
- Morgan Stanley Capital International and S&P Dow Jones. Revisions to the global industry classification standard (gics) structure, 2018.
- M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 2006.
- F. Nie, X. Wang, M. I. Jordan, and H. Huang. The constrained Laplacian rank algorithm for graph-based clustering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1969–1976, 2016.
- J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész. Dynamic asset trees and black monday. *Physica A: Statistical Mechanics and its Applications*, 324(1):247 – 252, 2003a.
- J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 2003b.
- J.-P. Onnela, K. Kaski, and J. Kertész. Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38:353–362, 2004.
- J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Society for Industrial and Applied Mathematics, 2000.
- S. Pal, F. Regol, and M. Coates. Bayesian graph convolutional neural networks using non-parametric graph learning. In *International Conference on Learning Representations (ICLR'19)*, 2019.
- N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.
- E. Pavez, H. E. Egilmez, and A. Ortega. Learning graphs with monotone topology properties and multiple connected components. *IEEE Transactions on Signal Processing*, 66(9): 2399–2413, 2018.
- V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Phys. Rev. Lett.*, 83:1471–1474, Aug 1999. doi: 10.1103/PhysRevLett.83.1471.
- V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley. Random matrix approach to cross correlations in financial data. *Physical Review E*, 65, Jun 2002.
- T. Raffinot. Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management*, 44, 2018a.
- T. Raffinot. The hierarchical equal risk contribution portfolio. *SSRN Electronic Journal*, 2018b.

- R. Ramakrishna, H. Wai, and A. Scaglione. A user guide to low-pass graph signal processing and its applications. *arXiv e-prints: 2008.01305*, 2020.
- S. I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer-Verlag New York, 2007.
- H. Rue and L. Held. *Gaussian Markov Random Fields: Theory And Applications*. Chapman & Hall/CRC, 2005.
- M. Różański, R. Wituła, and E. Hetmaniok. More subtle versions of the Hadamard inequality. *Linear Algebra and its Applications*, 532:500 – 511, 2017.
- A. Saha and A. Tewari. On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization*, 23(1):576–601, 2013.
- A. Schreiner. *Equity Valuation Using Multiples: An Empirical Investigation*. Springer, 2019.
- S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro. Network topology inference from spectral templates. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):467–483, 2017.
- R. Shafipour and G. Mateos. Online topology inference from streaming stationary graph signals with partial connectivity information. *Algorithms*, 13(9), July 2020.
- S. Shalev-Shwartz and A. Tewari. Stochastic methods for  $\ell_1$ -regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011.
- W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964.
- M. Slawski and M. Hein. Estimation of positive definite m-matrices and structure learning for attractive gaussian markov random fields. *Linear Algebra and its Applications*, 473: 145 – 179, 2015.
- J. A. Soloff, A. Guntuboyina, and M. I. Jordan. Covariance estimation with nonnegative partial correlations. *arXiv e-prints: 2007.15252*, July 2020.
- Standard & Poor’s. Global Industry Classification Standard (GICS). *Tech Report*, 2006.
- S. Sun, Y. Zhu, and J. Xu. Adaptive variable clustering in Gaussian graphical models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 931–939, 2014.
- Y. Sun, P. Babu, and D. P. Palomar. Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Transactions on Signal Processing*, 64(14): 3576–3590, 2016.
- Y. Sun, P. Babu, and D. P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2017.

- M. A. Sustik and B. Calderhead. Glassofast: An efficient glasso implementation. The University of Texas at Austin, UTCS Technical Report TR-12-29, 2012.
- K. M. Tan, D. Witten, and A. Shojaie. The cluster graphical Lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis*, 85:23 – 36, 2015.
- R. S. Tsay. *Analysis of Financial Time Series*. Wiley, 3rd edition, 2010.
- Y. Wald, N. Noy, G. Elidan, and A. Wiesel. Globally optimal learning for structured elliptical losses. In *Advances in Neural Information Processing Systems (NeurIPS'19)*, 2019.
- Y. Wang, U. Roy, and C. Uhler. Learning high-dimensional gaussian graphical models under total positivity without adjustment of tuning parameters. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 2698–2708, 2020.
- D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(3):615–636, 2009.
- D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.
- S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, June 2015.
- T. T. Wu and K. Lange. The MM alternative to EM. *Statistical Science*, 25(4):492–505, 2010.
- Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *arXiv e-prints: 1901.00596*, 2019.
- L. Yang, Y. Yang, G. B. Mgaya, B. Zhang, L. Chen, and H. Liu. Novel fast networking approaches mining underlying structures from investment big data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1–11, 2020.
- Y. Ye and E. Tse. An extension of karmarkar’s projective algorithm for convex quadratic programming. *Mathematical Programming*, 44:157—179, 1989.
- J. Ying, H. Lu, Q. Wei, J.-F. Cai, D. Guo, J. Wu, Z. Chen, and X. Qu. Hankel matrix nuclear norm regularized tensor completion for  $n$ -dimensional exponential signals. *IEEE Transactions on Signal Processing*, 65(14):3702–3717, 2017.
- J. Ying, J. V. de M. Cardoso, and D. P. Palomar. Does the  $\ell_1$ -norm Learn a Sparse Graph under Laplacian Constrained Graphical Models? *arXiv e-prints: 2006.14925*, June 2020a.
- J. Ying, J. V. de M. Cardoso, and D. P. Palomar. Nonconvex Sparse Graph Learning under Laplacian-structured Graphical Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.
- L. Zhao, Y. Wang, S. Kumar, and D. P. Palomar. Optimization algorithms for graph laplacian estimation via ADMM and MM. *IEEE Transactions on Signal Processing*, 67(16):4231–4244, 2019.