Information Sciences 608 (2022) 655–669

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Topological regularization with information filtering networks

Tomaso Aste

[a] Department of Computer Science, UCL, London, UK
[b] UCL Centre for Blockchain Technologies, UCL, London, UK
[c] Systemic Risk Centre, London School of Economics, London UK

ARTICLE INFO

Article history:
Received 14 September 2021
Received in revised form 24 March 2022
Accepted 4 June 2022
Available online 15 June 2022

Keywords:
Topological regularization
Information filtering networks
Complex systems
Covariance selection
Sparse inverse covariance
Chow-Liu Trees
Sparse expectation-maximization
IFN regression

ABSTRACT

This paper introduces a novel methodology to perform topological regularization in multivariate probabilistic modeling by using sparse, complex, networks which represent the system's dependency structure and are called information filtering networks (IFN). This methodology can be directly applied to covariance selection problem providing an instrument for sparse probabilistic modeling with both linear and non-linear multivariate probability distributions such as the elliptical and generalized hyperbolic families. It can also be directly implemented for topological regularization of multicollinear regression. In this paper, I describe in detail an application to sparse modeling with multivariate Student-t. A specific expectation–maximization likelihood maximization procedure over a sparse chordal network representation is proposed for this sparse Student-t case. Examples with real data from stock prices log-returns and from artificially generated data demonstrate applicability, performances, robustness and potentials of this methodology.

1. Introduction

Several real systems, such as the human brain or financial markets, are complex systems comprising a large number of interrelated variables. Data-driven, probabilistic modeling of these systems requires the estimate of the multivariate probability which, in general, depends on a large number of parameters. In order to reduce model complexity and increase model's generalization ability one would aim to sparsify the model by keeping only the set of most relevant and most reliable parameters. In some contexts, this simplification of the model through sparsification is a form of regularization. Regularization is an important tool in machine learning which is used to reduce the tendency of models to overfit the dataset on which they are trained and then underperform on new data. The common approach to regularization consists in adding a penalization term to the objective function in order to control for the complexity of the model with the aim of reducing overfitting. The idea was originally introduced by [1] and since then it has permeated the field of inverse problems and machine learning. The original Tikhonov approach (also known as ridge regression) was introduced in the context of multicollinear regression and consisted in penalizing the sum-of-square loss function by adding the sum of square of the regression coefficients (the L_2-norm) giving in this way preference to models with smaller coefficients. Other forms of penalization have been implemented and a particularly successful one uses of the L_1-norm, instead of the L_2-norm, and it was named 'lasso' by [2]. One of the consequences of the L_1-norm penalization is that, while some coefficient are shrank but rest finite, other

Abbreviations: TMFG, Triangulated Maximally 94 Filtered Graph; MFCF, Maximally Filtered Clique Forests; IFN, Information Filtering Networks; LASSO, Least Absolute Shrinkage and Selection Operator; Glasso, Graphical-lasso; EM, Expectation Maximization.

https://doi.org/10.1016/j.ins.2022.06.007
0020-0255/

are authomatically set to zero producing therefore sparse models. Penalization by the number of parameters is the L_0 norm which counts directly the number of non-zero coefficients and penalizes denser models. An advantage of the L_0-norm penalization is that the non-zero coefficients are not shrank in value allowing, in some cases, better optimizations. However, this has been proven to be a NP-hard problem being non-differentiable and having a combinatorically large number of possible configurations to be explored. There are different kinds of regularizations beyond the penalization based. Indeed, the regularization proposed in this paper does not use a penalization term but rather on a topological constraint. There are several regularization approaches that use geometry and topology and their combination (homology) to simplify models. For instance [3] proposed to account for the topological complexity of decision boundaries and penalize in favor of simpler models. Similarly, [4] analyzed the importance of topological elements devising a penalty function to retain the meaningful parts while discarding spurious topological structures. In the same line of approach, [5] penalizes complexity of neural networks decision boundaries by using persistent homology. (see, [6] for a broader perspective on this domain.).

In this paper, I propose a novel topological regularization approach over a sparse network representation. This approach applies directly to multivariate probabilistic modeling with the vast families of elliptical and generalized hyperbolic distributions. For these probability families, the dependency is encoded in a positively defined matrix **J** who's inverse (sometimes called scale matrix) is the covariance matrix, when the latter is defined. The present, proposed, regularization procedure forces to zero all elements of **J** that do not belong to the edge set of a specific, chordal, network representation. Maximum likelihood optimization over the network representation ensures that the sparse **J** is a proper scale matrix (positively defined). This topological regularization approach is in general applicable to the covariance selection problem and it can be therefore used for any modeling with a multivariate probability distribution with the covariance matrix in the parameter's set.

This approach can be formulated as a likelihood optimization under a topological constraint. In other words, the chordal network representation is a Bayesian prior for the inference model and the posterior probability is optimized given that prior structure. The network itself is learned maximizing likelihood with a Bayesian updating. In this paper, I will show that this problem can be solved for both multivariate normal and multivariate Student-t modeling. The multivariate normal case (see Section 2.4) is rather straightforward and it was already implicitly used in [7]. Conversely, the sparse model optimization for the Student-t case is less trivial. To the best of my knowledge, the proof of topologically-constrained likelihood optimization for Student-t models is an original result that I report in this paper for the first time (see Section 2.5).

The sparse chordal network representation that I use for the topological regularization is a chordal information filtering network with a structure that, while maximizing model's likelihood, yields to a meaningful network associated with the system's dependency structure. There is a vast, and long standing, literature addressing the problem of extracting such sparse network representations of the dependency structure [8–11]. These methodologies are often referred to as 'correlation networks' [11] or 'information filtering networks' (IFN) [9]. They distinguish slightly, with the correlation networks being normally obtained by retaining only the largest correlations, while IFN networks being constructed imposing topological constraints (such as being a tree or a planar graph) and optimizing some global properties (such as the likelihood). These network representations have been applied to a vast range of systems from finance [12–17], to psychology [18,19] and biology [20,21]. These network representations are quite powerful because they provide an intuitive description of the system yielding therefore to greater understanding of the system and a better interpretability of its models. However, until recent, these network representations have been mainly used as descriptive tools with limited quantitative analytics mainly associated to the relative position of the elements in the network, retrieving centrality measures or clustering properties. In [22] it was introduced a class of IFN, named Triangulated Maximally Filtered Graph (TMFG), that is chordal and therefore particularly suited for probabilistic inference modeling with sparse dependency structure. It was indeed shown in [7] that TMFG networks can be used, as Markov random fields, within a local–global procedure (LoGo) for probabilistic modeling, this was the first time a network from the IFN family was used to sparsify a multivariate probability model. In the multivariate normal case, the model inverse covariance is sparse with non-zero elements coinciding with the edges of the IFN structure. Sparse multivariate normal models produced with this procedure were proven to be extremely effective with larger likelihood performances than the graphical lasso (Glasso) [23] and with considerably lower computational burden. However, the Markov random field approach results, in general, in a complex expression for the multivariate probability which only for the multivariate normal distribution yields to a simpler multivariate normal model with sparse inverse covariance. In this paper, I discuss that for the vast families of elliptical and generalized hyperbolic multivariate probability distributions the sparse inverse covariance constructed over the IFN representation can be used as a form of topological regularization. This sparse inverse covariance, computed over the IFN structure, provides several advantages: (i) the model has greater interpretability, because the sparse structure is a meaningful representation of relevant interactions between the variables; (ii) the estimation of the model-parameters (i.e. the non-zero inverse covariance elements) is more accurate because only local inversions over significant correlations must be performed; (iii) the computational complexity is largely reduced because only a sub-set of parameters must be computed.

This work adds to the existing literature on two main original contributions. First, I apply for the first time information filtering networks to sparsify, multivariate probabilistic modeling with elliptical distributions, demonstrating that the spatrsification procedure is a form of topological regularization. Second, I originally extend the expectation maximization procedure to the parameter optimization for sparse Student-t. Furthermore, with extensive testing with real and simulated data, in this paper, I show that topological regularization with information filtering networks is a very effective tool providing models with better likelihoods, larger sparsity and better interpretability than state-of-the-art tools such as Glasso.

The paper is organize as follows. Section 2 presents the overall methodology including original proofs. Specifically, in subSection 2.1 I explain how an ensemble of data is generated by sub-sampling the real data and by synthetic generation. In subSection 2.2, I describe the construction of information filtering networks. The topological regularization approach is presented in subSection 2.3 where proof of parameter's optimization for the sparse topologically constrained model are provided. Examples of the application of this methodology to real and synthetic data are provided in Section 3. Conclusions and perspectives are provided in Section 4. Proof of theorems and methodological details are given in the appendixes.

2. Methodology

The overall methodology flowchart is depicted in Fig. 1. It consists in four main parts: a data pre-processing sub-sampling unit (used only for the real data); the information filtering network unit; the optimization unit and; the likelihood computation unit. Let me hereafter describe each of these parts separately.

2.1. Data ensemble generation

As input I use real log-return financial data ($\hat{x}_i(s) = \log Price_i(s) - \log Price_i(s-1)$) or synthetic data generated either with multivariate normal or multivariate Student-t. Real-data are sub-sampled from a dataset of daily prices from 623 stocks continuously traded on the US equity market between 01/02/1999 and 20/03/2020 for a total of 5515 trading days. For each stock 'i' $(=1,\dots,p)$. I performed sub-sampling both in time and on assets. Specifically, on the time-dimension, I randomly sample $2\times q$ days without repetition assigning q observations for the training set and q observations for the test set. In the experiments, I used $q=150$ and $q=600$. On the asset dimension, I piked randomly $p=100$ different return series among the 623 stocks. I repeated the operation 100 times generating therefore, for each q, an ensemble of 100 multivariate datasets with different sampled times and different compositions of assets.

Synthetic data were generated from multivariate normal distributions and multivariate Student-t distributions using the sample covariance and means from the real data as parameters to generate artificial datasets with properties consistent with the real data. The Student-t was generated with $\nu=2.2$ degrees of freedom. Analogously with the real data, I generated 100 random datasets of $p=100$ multivariate variables. I used $q=600$ observations for the training set and also $q=600$ observations for the test set. No pre-processing was used in this case.

2.2. Information filtering network learning

The structure of chordal IFN can be learned by using a clique expansion procedure where a clique forest is constructed starting from a seed structure and including vertices into the forest one by one accordingly with a given gain function (see [24,25] for further, deeper, insights on this approach). This is procedure is described in Algorithm 1. The resulting network is named MFCF (Maximally Filtered Clique Forests [18]). Such a clique forest network is made of a set of cliques $\mathscr{C}$ that are the 'vertices' in the clique-forest structure, the 'edges' of the clique-forest structure are instead a set $\mathscr{S}$ of separators that are cliques themselves with the property that by removing one of them the connected component becomes separated into two or more components. Clique forests are chordal graphs.

Algorithm 1: The MFCF clique forest construction.

input A gain function $G(\cdot,\cdot)$.
input the minimum clique size Min_Cl.
input the maximum clique size Max_Cl.
input the number of times separators can be used (multiplicity) Max_Mult.
initialize Start from a seed clique $\mathbf{c}_0$ with vertices $\mathbf{v}_0$ and edges $\mathbf{e}_0$.
initialize $V \leftarrow \mathbf{v}_0$, $E \leftarrow \mathbf{e}_0$, $\mathcal{C} \leftarrow \mathbf{c}_0$.
while *There are still vertices not included in the MFCF,* $|V| < p$ **do**

- Find a vertex, $v \notin V$, and a sub-clique, $\mathbf{s}$, with size larger than Min_C -1 but smaller than Max_Cl, that return the largest gain: $v = \max_{k\notin V}(G(k,\mathbf{s})|\mathbf{s}\subseteq\mathbf{c}\in\mathcal{C} \text{ and Min\_C} - 1 \le |\mathbf{s}| < \text{Max\_Cl})$.
- Create a new clique $\mathbf{c}' = \mathbf{s}\cup v$
- Add the new clique $\mathbf{c}'$ to the clique set $\mathcal{C}\leftarrow\mathbf{c}'$.
- If $\mathbf{s}=\mathbf{c}$ remove the clique $\mathbf{c}$ from $\mathcal{C}$.
- Update $V\leftarrow v$, $E\leftarrow$ *(edges between* v *and* $\mathbf{c}$*).*

output The MFCF(Min_Cl,Max_Cl,Max_Mult): $\mathcal{G}=(V,E)$.

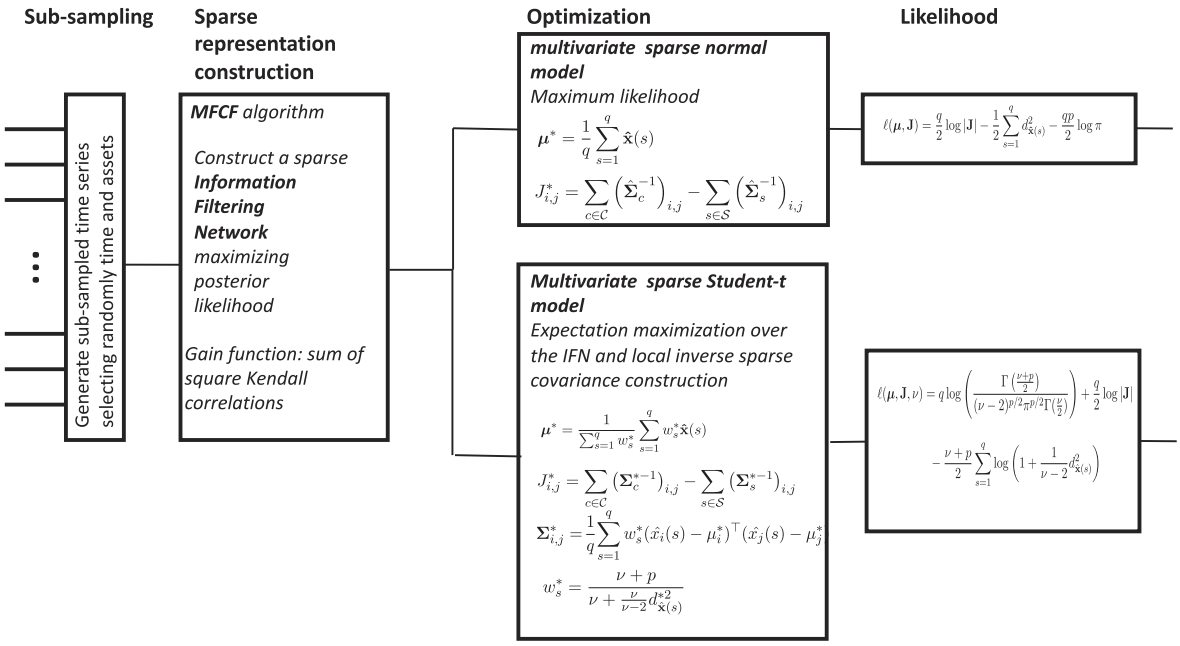

Fig. 1. Schematic representation of the overall framework of the process presented in this paper. It consists in four main parts: a data pre-processing unit, where sub sampled time-series are generated from the real financial data; the information filtering network unit, where the sparse network representation is generated; the optimization unit, where optimal parameters are estimated and; the likelihood computation unit, where likelihoods are computed. Both processes for multivariate normal and multivariate Student-t are represented. For artificial data, the data pre-processing unit is not used.

The MFCF network complexity can be constrained by limiting the minimum and maximum clique sizes (Min_Cl and Max_Cl parameters in Algorithm 1). By increasing the clique sizes one increases the number of edges in the network making it denser. The full network is retrieved when the minimum clique size equals the total number of vertices. Separators can be constrained to be unique between two cliques (multiplicity one) or to be utilizable more than once by more than two cliques (multiplicity larger than one, see Max_Mult parameter in Algorithm 1). The simplest clique is the 2-clique that has two elements and it is an edge. MFCF networks with two cliques only are segments if separators have multiplicity one, or they are maximum spanning threes when separators have arbitrary multiplicity. The TMFG, first introduced in [22], is a MFCF obtained when cliques have all size 4 (tetrahedra) and the separators can be used only once. The networks that I use in this paper have minimum clique size equal to 2 and a maximum clique sizes ranging between 2 to the total number of vertices, and the multiplicity is fixed to 1.

Algorithm 1 requires a gain function that is used to decide the inclusion of a vertex into the clique forest in a recursive way. The choice of a convenient gain function is strictly related to the problem under investigation. The gain function that I use in this paper is the sum of the squares of the coefficients of the correlation matrix. This gain function is a good proxy for likelihood in a range of problems. This is a very simple gain function that lead to networks with all cliques with maximum size. Indeed, with this kind of additive gain the algorithm always gains by enlarging the clique, if allowed. I test both Pearson and Kendall correlations. The latter describe dependency for a broader class of multivariate random variables than the Pearson's correlations, they are non-linear and have been proven to be effective in practical applications [14].

The IFN structure is learned before the estimate of the model-parameters and it is passed to the optimization procedure as a Bayesian prior. This approach is analogous to the LoGo methodology introduced in [7]. However, here we apply it to non-normal models and this has important implications. Indeed, outside normal modeling the structure of the IFN graph does no longer represent conditional independence and the sparse probability distribution function no longer factorizes over the IFN clique and separator structure (see [26] and Eq. 12 in Appendix A for this factorization for the multivariate normal probability distribution function case).

2.3. Optimization with topological regularization

The optimization problem consists in finding the model parameters that maximize likelihood for a given IFN. In this paper, I address this issue for probabilistic modeling with densities belonging to the elliptical family and I report results for the multivariate normal and Student-t cases.

For the whole elliptical family the probability density function can be written as:

$$f(\mathbf{X}=\mathbf{x}) = k_p\sqrt{|\mathbf{J}|}g(\mathbf{x},\boldsymbol{\mu},\mathbf{J}), \tag{1}$$

where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ are the expected values of $\mathbf{X}$ and $\mathbf{J}$ is a positively defined matrix which coincides with the inverse covariance matrix when it is defined [27].

The matrix $\mathbf{J}$ is the quantity I am sparsifying in this proposed topological regularization procedure. Specifically, only the diagonal $(\mathbf{J})_{i,i}$ and elements $(\mathbf{J})_{i,j}$ corresponding to edges in the IFN are allowed to be different from zero. Therefore, the optimization problem becomes computing the values of the non-zero elements of $\mathbf{J}$ which maximize the likelihood. Hereafter, I report the solution for the multivariate normal and Student-t cases.

2.4. Sparse maximum likelihood solution for the multivariate normal

The log-likelihood for the multivariate normal distribution is.

Definition 1 (*Normal log-likelihood*). Given a set of observations $\hat{\mathbf{x}}(s) = (\hat{x}_1(s),\dots,\hat{x}_p(s))^{\mathsf{T}}$ with $s = 1\dots q$, the log-likelihood of the multivariate normal is

$$\ell(\boldsymbol{\mu},\mathbf{J}) = \frac{q}{2}\log|\mathbf{J}| - \frac{1}{2}\sum_{s=1}^{q} d^2_{\hat{\mathbf{x}}(s)} - \frac{qp}{2}\log\pi. \tag{2}$$

where

Definition 2 (*Mahalanobis distance*). The term

$$d^2_{\hat{\mathbf{x}}(s)} = (\hat{\mathbf{x}}(s)-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{J}(\hat{\mathbf{x}}(s)-\boldsymbol{\mu}), \tag{3}$$

is the square of the Mahalanobis distance [28].

Remark 1. The matrix $\mathbf{J}$ in Eqs. 2 and 3 must be positively defined and sparse with non-zero elements only allowed on the diagonal and or in the off-diagonal positions coinciding with the edges of the associated IFN.

Now I must find the maximum likelihood solution of Eq. 2 under the topological constraint that off-diagonal non-zero elements of $\mathbf{J}$ must coincide with the edges of the given IFN. The maximization process is almost identical to the full case but with the topological constraint enforced.

Theorem 1 (*ML solution for $\boldsymbol{\mu}$ for the sparse multivariate normal problem*). *If $\mathbf{J}$ is invertible, then the maximum likelihood solution for $\boldsymbol{\mu}$ is the sample mean:*

$$\boldsymbol{\mu}^* = \frac{1}{q}\sum_{s=1}^{q}\hat{\mathbf{x}}(s) \tag{4}$$

Proof. The proof is identical to the one for the full problem. The maximum of $\ell(\boldsymbol{\mu},\mathbf{J})$ in Eq. 2 with respect to $\boldsymbol{\mu}$ is obtained from the root of

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ell(\boldsymbol{\mu},\mathbf{J}) = \frac{1}{2}\mathbf{J}\sum_{s=1}^{q}\hat{\mathbf{x}}(s) - \frac{q}{2}\mathbf{J}\boldsymbol{\mu} = 0. \tag{5}$$

Which is indeed solved by $\boldsymbol{\mu}^*$ if $\mathbf{J}$ is invertible. □

The proof that $\mathbf{J}$ is invertible is given in Lemma 1 in Appendix B.

Theorem 2 (*ML solution for $\mathbf{J}$ for the sparse multivariate normal problem*). *Given a IFN structure made of clique and separators, the maximum likelihood solution for the sparse $\mathbf{J}$ is:*

$$J^*_{i,j} = \sum_{c\in\mathscr{C}}(\hat{\boldsymbol{\Sigma}}_c^{-1})_{i,j} - \sum_{s\in\mathscr{S}}(\hat{\boldsymbol{\Sigma}}_s^{-1})_{i,j}, \tag{6}$$

when i,j belong to a clique of the IFN. Otherwise $J^*_{i,j} = 0$ for all other couples of i,j not belonging to cliques. Where $\hat{\boldsymbol{\Sigma}}_c$ and $\hat{\boldsymbol{\Sigma}}_s$ are the Person's sample estimators of the covariances of the variables in the cliques and separators.

The proof of this theorem is provided in Appendix B.

Remark 2. The sparsification of **J** through Eq. 6 provides a way to overcome the curse of dimensionality in the estimation of covariances from observations. Indeed, independently on the overall dimension of the system of variables **X**. When the sparse inverse covariance **J** is estimated from data, it is then sufficient to have a number of observations, q, larger than the size of the largest clique, which is independent from the dimension, p, of **X**. Therefore, through Eq. 6 one can obtain well conditioned covariance matrices even when $q \ll p$. Eq. 6 transforms the global problem of estimating the whole matrix inverse into a set of local problems at clique and separator levels.

2.5. Sparse maximum likelihood solution for the multivariate Student-t

Definition 3 (*Student-t log-likelihood*). Given a set of observations $\hat{\mathbf{x}}(s) = (\hat{x}_1(s), \dots, \hat{x}_p(s))^\top$ with $s = 1 \dots q$, the log-likelihood of the multivariate Student-t is

$$\ell(\boldsymbol{\mu}, \mathbf{J}, \nu) = q \log\left(\frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu-2)^{p/2}\pi^{p/2}\Gamma\left(\frac{\nu}{2}\right)}\right) + \frac{q}{2}\log|\mathbf{J}| - \frac{\nu+p}{2}\sum_{s=1}^{q}\log\left(1 + \frac{1}{\nu-2}d^2_{\hat{\mathbf{x}}(s)}\right), \tag{7}$$

for $\nu > 2$.

where $d^2_{\hat{\mathbf{x}}(s)}$ is the square Mahalanobis distance as defined in Definition 2; **J** is the inverse covariance and ν is the degrees of freedom that here we assume being always larger than 2. Indeed, for $\nu \leqslant 2$ the covariance is not defined.

In the non-sparse (full) case, it is known that the likelihood of multivariate Student-t models can be maximized by means of a procedure known as expectation–maximization (EM) introduced by [29] (see also [30], Chap.9).

I shall show hereafter that such a procedure can be applied also to the maximization the likelihood of the sparse Student-t model for any given chordal IFN structure.

Theorem 3 (*ML solution for* $\boldsymbol{\mu}$ *for the sparse multivariate Student-t problem*). *If* **J** *is invertible, then the maximum likelihood solution for* $\boldsymbol{\mu}$ *is a weighted mean:*

$$\boldsymbol{\mu}^* = \frac{1}{\sum_{s=1}^{q} w_s^*}\sum_{s=1}^{q} w_s^* \hat{\mathbf{x}}(s), \tag{8}$$

with weights

$$w_s^* = \frac{\nu+p}{\nu + \frac{\nu}{\nu-2}d^{*2}_{\hat{\mathbf{x}}(s)}}. \tag{9}$$

Proof is provided in Appendix C.

The $d^{*2}_{\hat{\mathbf{x}}(s)}$ is the Mahalanobis distance computed using the ML solution $\mathbf{J}^*$ (see Theorem 4). The w_s^* are the asymptotic solutions for $t \to \infty$ of the recursive EM process.

The ML solution for the sparse **J** is also obtained with the EM approach.

Theorem 4 (*ML solution for* **J** *for the sparse multivariate Student-t problem*). *The maximum likelihood solution for* **J** *is:*

$$J^*_{i,j} = \sum_{c\in\mathscr{C}}\left(\boldsymbol{\Sigma}_c^{*-1}\right)_{i,j} - \sum_{s\in\mathscr{S}}\left(\boldsymbol{\Sigma}_s^{*-1}\right)_{i,j}, \tag{10}$$

when i,j *are an edge of a clique. Otherwise* $J^*_{i,j} = 0$ *for all other couples of* i,j *not belonging to cliques. Where* $\boldsymbol{\Sigma}_c^*$ *and* $\boldsymbol{\Sigma}_s^*$ *are the EM estimators of the covariances of the variables in the cliques and separators, which are given by the weighted sample averages:*

$$\boldsymbol{\Sigma}^*_{i,j} = \frac{1}{q}\sum_{s=1}^{q} w_s^*(\hat{x}_i(s) - \mu_i^*)^\top(\hat{x}_j(s) - \mu_j^*). \tag{11}$$

The proof of this theorem is provided in Appendix C.

The sparsity of **J** is not affecting the form of the EM solutions which have the same form also in the full case. However, in the sparse case only the elements belonging to cliques must be computed which reduces computational complexity from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$.

The parameter ν can also be computed through the EM procedure. However, I prefer to estimate it independently via a power law fit of the left and right tails of the probability distribution of all the univariate marginals of **X**. Indeed, all marginal Student-t distributions of **X** must behave as power laws on both left and right tails with tail-exponent ν.

3. Experiments

In order to test the novel topological regularization methodology introduced with this paper I computed the likelihood of several models using three kind of data. I then compared the results obtained with IFN sparsification with results obtained with -state-of-the-art- graphical lasso sparsification.

3.1. Model construction and parameter estimation

For each training dataset, I generated MFCF networks with maximum clique sizes in the range from 2 to 100. For each maximum clique size, I generate two different networks by using the Pearson correlation estimate and the Kendall correlation estimate. As MFCF gain function I chose the sum of the squares correlations, which is one of the simplest choices that produces cliques all of sizes equal to the maximum clique size. The MFCF networks I generate have separators that are used only once (multiplicity one). As degrees of freedom I empirically investigated the tails of the marginal distributions across the whole dataset retrieving a tail exponent $v = 2.2$ as a good average estimator for the degrees of freedom. I verified that relative results are little sensitive to this parameter although the values the likelihood can change sensibly with v. The covariances are retrieved by multiplying by the standard deviations the elements of the correlations matrices. Using these MFCF networks I then compute the maximum likelihood inverse sparse covariance estimates for multivariate normal modeling, as described in Eq. 6, and for the multivariate Student-t modeling, as described in Eq. 10.

I summary, I test six combinations of models: 1. multivariate normal with Pearson correlations (Nor.; Per.); 2. multivariate normal with Kendall correlations (Nor.; Ken.); 3. multivariate Student-t with Pearson correlations (St-t.; Per.); 4. multivariate Student-t with Pearson correlations optimized with Expectation Maximization (St-t.; Per. EM), 5. multivariate Student-t with Kendall correlations (St-t.; Ken.); 6. multivariate Student-t with Kendall correlations optimized with Expectation Maximization (St-t.; Ken. EM).

3.2. Comparison with Glasso

In order to compare the results with a meaningful state-of-the-art sparse modeling approach, I computed L_1-norm regularized sparse inverse covariance estimators by using a Quadratic Approximation for Sparse Inverse Covariance Estimation (QUIC) by [31].[1] Different levels of sparsity were achieved by varying the regularization penalty, λ, with values between 10^{-6} and 10^{-3}.

3.3. Results

I computed the mean log-likelihood ℓ for the range of MFCFs with different clique sizes and for both multivariate normal and multivariate Student-t models computed by using either the Pearson's and the Kendall's covariance estimators and the Expectation Maximization procedure (six models in total, see Section 2.5). The largest mean log-likelihoods across the MFCF clique-sizes' range and the value of the corresponding clique size are reported in Table 1 for all the models. The parameters are estimated on the training set and the results are instead reported for the test set. One can observe that for real data the sparse Student-t model constructed by using Kendall's covariance and Expectation Maximization procedure gives the best results for both $q = 150$ and 600 with smaller clique size selected for the shorter dataset. The combination Student-t model, Kendall's covariance and Expectation Maximization procedure is also best for the multivariate Student-t synthetic datasets. Conversely, for the multivariate Normal synthetic datasets the best results are achieved by the sparse Normal model construct using Peterson's covariance.

Fig. 2 reports results for the Student-t log-likelihood (Eq. 7) estimated using Kendall covariance and expectation maximization. The parameters are estimated on the training set and the results are instead reported for the test set. The log-likelihood is computed for a range of sparsity values obtained by varying the maximum clique size from 2 to 100 (the second being the complete graph). The x-axis, $\|\mathbf{J}\|_0$, reports the number of edges in MFCF (i.e. the number of non-zero off-diagonal elements in the sparse inverse covariance). The tick lines are averages over 100 re-samplings and the bands are the 10% and 90% quantiles. Note that, the re-sampling picks randomly both the time series and the returns. Therefore the observed consistency and the relatively narrow quantile band are strong indications of statistical robustness of the results. Also note that the last points on the right of the two plots are the full models (max clique = 100) with the complete (non-sparse) inverse covariance matrix. As one can see, for small observation sets ($q = 150$) the sparse models greatly over-perform the complete models. Whereas, for larger observation sets ($q = 600$) the difference is smaller.

In the figure, I also report for comparison results obtained by estimating sparse inverse covariance via L_1-norm regularization using Glasso with the QUIC package by [31]. One can see that for sparse modeling, up to $\|\mathbf{J}\|_0 \sim 10 \times p = 1,000$ the MFCF approach is largely over-performing the QUIC-Glasso results. For denser networks (i.e. $\|\mathbf{J}\|_0 \sim 2,000$) the MFCF and QUIC approach deliver similar results. For instance, for $q = 600$, the Glasso approach with $\lambda = 2 \cdot 10^{-5}$ retrieves 1,720 average number of edges and an average $\ell/q = 253.9$ with $[248.7, 258.6]$ the 10% and 90% quantiles. By comparison, the MFCF for max

[1] Matlab implementation available at: http://www.cs.utexas.edu/ sustik/QUIC/.

Table 1
Summary of test set results for the mean log-likelihood per observation (ℓ/q) for the various combinations of models (rows) and data types (columns). For each model and dataset type, the table reports only the best likelihood values obtained across the range of max-clique-sizes in the MFCF network. The ℓ/q values are on the left side of the columns while the corresponding max-clique-size are reported on the right side of the columns. Several modes are investigated: Normal likelihood (Nor., see Eq. 2), Student-t likelihood (Nor., see Eq. 7), Pearson covariance estimator (Per.), Kendall covariance estimator (Ken.), expectation maximization parameters estimation (EM, see Eqs. 8, 10).

Max average log-likelihood per observations ℓ/q; and max clique size								
Estimator	Real q = 150		Real q = 600		St-t. q = 600		Nor. q = 600	
Nor.; Per.	350.9;	5	362.6;	10	344.0;	5	**371.3;**	30
Nor.; Ken.	360.4;	20	363.8;	100	360.6;	100	369.4;	100
St-t.; Per.	376.7;	6	383.3;	11	447.0;	7	363.3;	30
St-t.; Per. EM	383.9;	8	389.6;	20	459.6;	50	366.8;	30
St-t.; Ken.	381.0;	15	385.6;	50	454.7;	100	364.8;	100
St-t.; Ken. EM	**384.8;**	8	**389.8**;	15	**460.0;**	30	366.8;	30

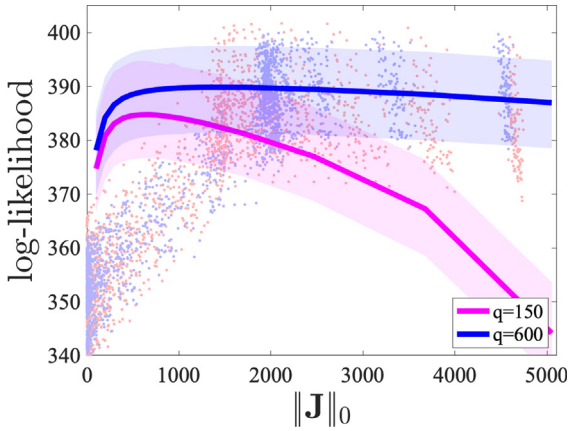

Fig. 2. Log-likelihood $\ell(\boldsymbol{\mu}, \mathbf{J}, v)/q$ (Eq. 7) for Student-t models with sparse inverse covariance matrix **J** constructed by using the MFCF approach with IFN graphs with different levels of sparsity obtained by changing the maximum clique size from 2 to 100. The IFN have been constructed using the Kendall estimate of the correlation matrix. The x-axis reports $\|\mathbf{J}\|_0$ which is the number of edges in the IFN graph. Models have been optimized to maximize Student-t likelihood by using the EM procedure described in Section 2.5. Parameters are estimated on two train sets of 'real' data (see text) with different lengths: $q = 150$ and $q = 600$ respectively (blue and magenta). Reported results are the log-likelihoods computed on the test set. The lines are the means and the bands around the lines are the 10% and 90% quantiles over 100 random re-sampling. The points are instead from Glasso models computed using QUIC Quadratic Approximation for Sparse Inverse Covariance Estimation implemented by [31] using a range of regularization penalty between 10^{-6} to 10^{-3}.

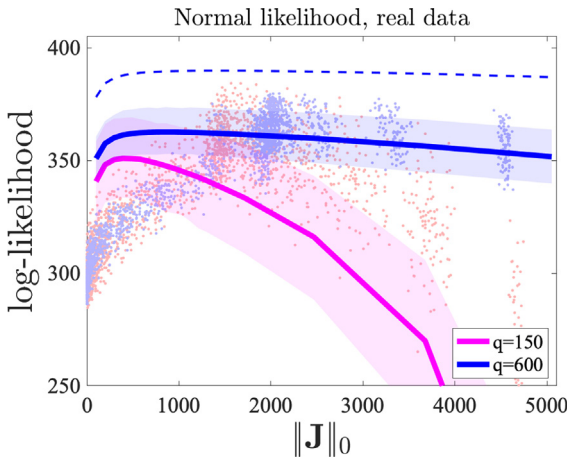

Fig. 3. Normal log-likelihood $\ell(\boldsymbol{\mu}, \mathbf{J}, v)/q$ (Eq. 2) for real financial data (see text) for a set of models with sparse inverse covariance matrix **J** constructed by using the MFCF approach with IFN graphs with different levels of sparsity obtained by changing the maximum clique size from 2 to 100. The IFN have been constructed using the Pearson estimate of the correlation matrix. The x-axis reports $\|\mathbf{J}\|_0$ which is the number of edges in the IFN graph. Parameters are estimated on two train sets with different lengths: $q = 150$ and $q = 600$ respectively (blue and magenta). Reported results are the log-likelihoods computed on the test set. The lines are the means and the bands around the lines are the 10% and 90% quantiles over 100 random re-sampling. The points are instead normal log-likelihood for Glasso models computed using QUIC using a range of regularization penalty between 10^{-6} to 10^{-3}. The slashed blue line is the Student-t likelihood for $q = 600$ reported in Fig. 2 which is reported for comparison.

clique equal to 20 has 1,710 edges and average likelihood $\ell/q = 254.0$ with quantiles $[248.9, 258.7]$. Similar results are retrieved for other levels of sparsity with the Glasso results slightly improving performances over the MFCF when the model becomes denser.

Other results with different combination of models and with artificial data are reported in Appendix D. Specifically, I test normal modeling on the real datasets (Fig. 3 and I test both normal and Student-t models on synthetic datasets produced with multivariate normal and Student-t distributions (Figs. 4 and 5). Overall, I observe a very consistent picture across al experiments and the various combinations of model constructions and data. It results that Student-t modeling is more appropriate for the real financial data resulting in larger likelihoods. This is consistent with current literature. Not surprisingly, it results that normal models works better on normal data and instead Student-t models have higher likelihoods on Student-t data. The construction with Kendall's estimate of the covariance is producing better results for the real data and the Student-t synthetic data but not for the multivariate normal synthetic data where the Pearson's estimate is better.

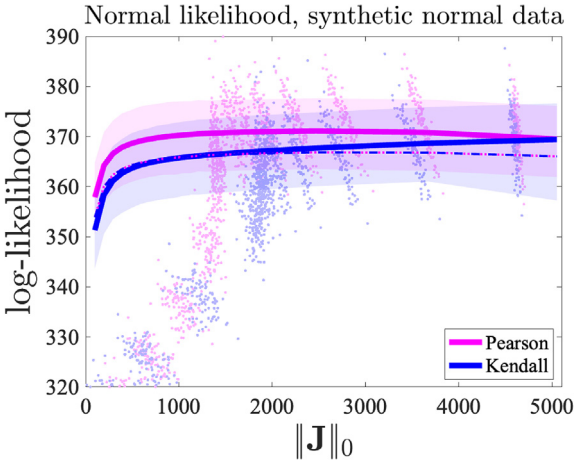

Fig. 4. Normal log-likelihood $\ell(\boldsymbol{\mu}, \mathbf{J}, v)/q$ (Eq. 2) for multivariate normal synthetic data (see text) for a set of models with sparse inverse covariance matrix $\mathbf{J}$ constructed by using the MFCF approach with IFN graphs with different levels of sparsity obtained by changing the maximum clique size from 2 to 100. The IFN have been constructed using both the Pearson and the Kendall estimates of the correlation matrix (magenta and blue lines respectively). The x-axis reports $\|\mathbf{J}\|_0$ which is the number of edges in the IFN graph. Parameters are estimated on train sets of lengths $q = 600$. Reported results are the log-likelihoods computed on the test set. The lines are the means and the bands around the lines are the 10% and 90% quantiles over 100 random re-sampling. The points are instead normal log-likelihood for Glasso models computed using QUIC using a range of regularization penalty between 10^{-6} to 10^{-3}. The slashed blue line and the dotted magenta line are the Student-t models with Kendall and Pearson estimates respectively; they overlap but do not coincide.

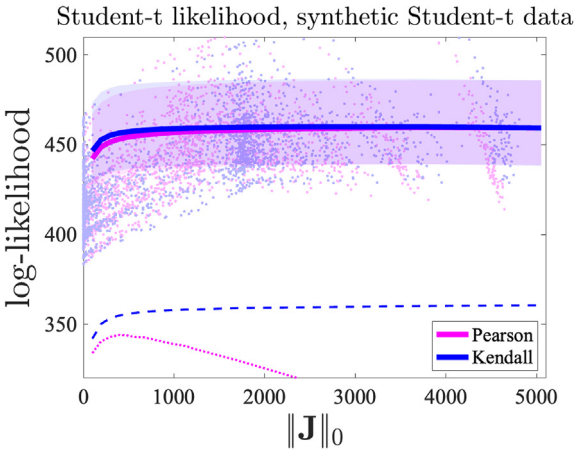

Fig. 5. Student-t log-likelihood $\ell(\boldsymbol{\mu}, \mathbf{J}, v)/q$ (Eq. 7) for multivariate Student-t synthetic data (see text) for a set of models with sparse inverse covariance matrix $\mathbf{J}$ constructed by using the MFCF approach with IFN graphs with different levels of sparsity obtained by changing the maximum clique size from 2 to 100. The IFN have been constructed using both the Pearson and the Kendall estimates of the correlation matrix (magenta and blue lines respectively). The x-axis reports $\|\mathbf{J}\|_0$ which is the number of edges in the IFN graph. Parameters are estimated on train sets of lengths $q = 600$. Reported results are the log-likelihoods computed on the test set. The lines are the means and the bands around the lines are the 10% and 90% quantiles over 100 random re-sampling. The points are instead normal log-likelihood for Glasso models computed using QUIC using a range of regularization penalty between 10^{-6} to 10^{-3}. The slashed blue line and the dotted magenta line are the normal models with Kendall and Pearson estimates respectively.

The expectation–maximization optimization procedure used for the Student-t models makes the difference between Kendall's and Pearson's estimates small, but still quantifiable. Indeed, EM can 'cure' the parameters estimate and therefore it is little sensitive to the starting matrix, however the IFM networks from Kendall's or Pearson's estimates are not identical and this produces the difference. Consistently with what I reported for the real data with Student-t modeling (Fig. 2), Glasso models underperform for all sparse networks and then achieve comparable performances to the IFM-LoGo models at higher levels of network density (above $\|\mathbf{J}\|_0 \sim 2,000$) which correspond to rather dense networks with about 40% of edges present.

4. Conclusions and perspectives

In this paper I have introduced a methodology for topological regularization with information filtering networks. I have demonstrated the application of this method for multivariate normal and multivariate Student-t sparse modeling using MFCF clique-forest networks. The regularization methodology consists in keeping different from zero only the parameters of the multivariate distribution that correspond to edges in the IFN. By using clique forests IFNs, one guarantees positive definiteness and decomposition into local parts of the inverse covariance matrix (Eqs. 6 and 10). This is an important property associated with this kind of IFNs and it applies to the vast class of models belonging to the elliptical family [27] which includes the Student-t but also the normal, the Laplace and the multivariate stable distributions.

The present topological regularization with IFN is actually more general, applying also to non symmetric multivariate distributions such as the generalized hyperbolic family and broadly to any multivariate modeling which makes use of the covariance matrix in the parameter set. This topological regularization methodology strongly improves model interpretability because IFN structures are known to meaningfully represent relevant interrelations in complex data structures with a vast literature reporting their successful descriptive power with applications to various domains from finance to biology.

I have shown that the expectation–maximization methodology commonly used to estimate the maximum likelihood coefficients in Student-t models can be used also for this regularized sparse models with the advantage that, in this case, computation must be done only for the coefficients corresponding to IFN edges reducing computation complexity from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$.

Experiments on real datasets from equity prices and multivariate synthetic datasets demonstrate that the proposed methodology is directly applicable to a range of practically relevant problems. Results demonstrate that topologically regularized models outperform full models and reveal that smaller observation sets are optimized by using sparser IFN models. A comparison with L_1-norm regularization by Glasso approach, shows that the proposed methodology largely outperforms Glasso for sparse models and tend to perform similarly to Galsso for denser models. Furthermore, if must be noticed that the proposed IFN-LoGo approach is computationally more efficient than Glasso and the sparse network has better interpretability.

The present paper reports exclusively on topological regularization via IFN priors, however, the nature of this sparsification allows to combine straightforwardly L_1 and L_2 regularizations within this methodology. Indeed, Theorems 2 and 4 provide a formula for the maximum likelihood solution of the sparse inverse covariance matrix as sum of local inverse matrices associated with the clique and separator sets. On such local inversions, shrinkage and lasso regularization can be applied directly. This has the further advantage that both the inversions and the regularizations are on local-small dimensional matrices making the procedure computationally efficient and fully parallelizable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author acknowledges discussions with several members of the Financial Computing and Analytics group at UCL. A special thank to Dr. Guido Massara for many critical inputs and to Killian Guillaume Paul Martin-Horgassan for discussions. Also, thanks for support from ESRC (ES/K002309/1), EPSRC (EP/P031730/1) and EC (H2020-ICT-2018–2 825215).

Appendix A. Decompositions for the multivariate normal case

Theorem 5 (*Decomposition of the sparse multivariate normal distribution*). *Given a sparse* $\mathbf{J}$ *inverse covariance with a chordal IFN structure where the off diagonal non-zero entries corresponds to a set of cliques* $\mathscr{C}$ *and separators* $\mathscr{S}$ *in a clique-forest, the sparse multivariate normal probability density function,* $\varphi(\mathbf{X} = \mathbf{x}|\boldsymbol{\mu}, \Sigma)$*, can be decomposed in terms of cliques and separators as follows:*

$$\varphi(\mathbf{X}=\mathbf{x}|\boldsymbol{\mu},\Sigma)=\frac{\prod_{c\in\mathscr{C}}\varphi(\mathbf{X_c}=\mathbf{x_c}|\boldsymbol{\mu}_c,\Sigma_\mathbf{c})}{\prod_{s\in\mathscr{S}}\varphi(\mathbf{X_s}=\mathbf{x_s}|\boldsymbol{\mu}_s,\Sigma_\mathbf{s})}. \tag{12}$$

Proof. The proof is a straightforward consequence of the exponential form of the normal distribution and it is for instance provided in [26]. □

Theorem 6 (*Decomposition of conditionally independent multivariate normal variables*). *Given a set of multivariate normal variables corresponding to a set of cliques $\mathscr{C}$ which are conditionally independent from each other when conditioned to their separators $\mathscr{S}$ in a clique-forest structure, then the multivariate normal probability density function can be decomposed in terms of cliques and separators as follows:*

$$\varphi(\mathbf{X}=\mathbf{x}|\boldsymbol{\mu},\Sigma)=\frac{\prod_{c\in\mathscr{C}}\varphi(\mathbf{X_c}=\mathbf{x_c}|\boldsymbol{\mu}_c,\Sigma_\mathbf{c})}{\prod_{s\in\mathscr{S}}\varphi(\mathbf{X_s}=\mathbf{x_s}|\boldsymbol{\mu}_s,\Sigma_\mathbf{s})}. \tag{12}$$

Proof. The proof is a direct consequence of the Bayes formula it is, for instance, provided in [26]. □

It is clear that the two formulas in Theorems 5 and 6 are identical (indeed they have been labeled with the same number, 12), however they are consequences of two different facts that happen to coincide for the multivariate normal probability density function.

Remark 3. The conditional independence is an exclusive property of the sparse multivariate normal and it is not applicable to the Student-t case.

As a consequence of Eq. 12 one has that the non-zero elements of the sparse covariance matrix **J** can be expressed as a simple sum of local inverse covariances.

Corollary 1 (*Decomposition of the inverse covariance matrix*). *The elemets of the inverse covariance are given by:*

$$J_{i,j}=\sum_{c\in\mathscr{C}}\left(\boldsymbol{\Sigma}_c^{-1}\right)_{i,j}-\sum_{s\in\mathscr{S}}\left(\boldsymbol{\Sigma}_s^{-1}\right)_{i,j}, \tag{13}$$

Proof. This is a direct consequence of Eq. 12 and the proof is provided in [26]. □

There are other two useful consequences of the decomposition in Eq. 12.

Corollary 2 (*Decomposition of the determinant*).

$$|\mathbf{J}|=\frac{\prod_{c\in\mathscr{C}}|\mathbf{J}_c|}{\prod_{s\in\mathscr{S}}|\mathbf{J}_s|}. \tag{14}$$

Proof. This is a direct consequence of Eq. 12 and the proof is provided in [26]. □

Corollary 3 (*Decomposition of the Mahalanobis distance*).

$$d^2=\sum_{c\in\mathscr{C}}d_c^2-\sum_{s\in\mathscr{S}}d_s^2 \tag{15}$$

with $d_c^2=(\mathbf{x}-\boldsymbol{\mu}_c)^\top\mathbf{J}_c(\mathbf{x}-\boldsymbol{\mu}_c)$ and $d_s^2=(\mathbf{x}-\boldsymbol{\mu}_s)^\top\mathbf{J}_s(\mathbf{x}-\boldsymbol{\mu}_s)$.

Proof. This is a direct consequence of Eq. 12 and the proof is provided in [26]. □

Appendix B. Theorems and proofs for normal ML

Lemma 1 (*Positive definitness*). *If all Σ_c with $c \in \mathscr{C}$ and Σ_s with $s \in \mathscr{S}$ are positively defined, the sparse inverse covariance* $\mathbf{J}$ *constructed from Eq. 6 is positively defined.*

Proof. From Eq. 14, if all $|\mathbf{J}_c| > 0$ with $c \in \mathscr{C}$ and $|\mathbf{J}_s| > 0$ with $s \in \mathscr{S}$, then $|\mathbf{J}| > 0$. □

Proof. (**Proof of Theorem 2**).

I have to prove that the sparse inverse covariance matrix constructed using Eq. 6 is the maximum likelihood solution for the sparse multivariate normal case for a give IFN sparsity structure.

I develop this proof into two steps.

1. **First**, I show that, if i,j is an edge of a clique of the IFN structure, then the solution for the covariance coefficient must be the Person's sample covariance estimator between variable i and variable j. This part proceed in the same way as for the full problem. In particular, the maximum of $\ell(\boldsymbol{\mu},\mathbf{J})$ with respect to $\mathbf{J}$ is obtained from the root of
$$\frac{\partial}{\partial J_{ij}}\ell(\boldsymbol{\mu},\mathbf{J}) = \frac{q}{2}(\mathbf{J}^{-1})_{i,j} - \frac{1}{2}\sum_{s=1}^{q}(\hat{x}_i(s)-\mu_i)(\hat{x}_j(s)-\mu_j)\Big|_{\mathbf{J}=\mathbf{J}^*} = 0, \tag{16}$$
for $(i,j) \in c$. This therefore implies that the elements i,j in the maximum likelihood covariance must coincide with the Person's sample covariance estimator, $(\hat{\boldsymbol{\Sigma}})_{i,j}$, when the couple i,j is an edge of a clique.
2. **Second**, I demonstrate that the sparsity structure of $\mathbf{J}$ over a chordal graph imposes that the inverse covariance must be in the form
$$J_{i,j} = \sum_{c\in\mathscr{C}}(\boldsymbol{\Sigma}_c^{-1})_{i,j} - \sum_{s\in\mathscr{S}}(\boldsymbol{\Sigma}_s^{-1})_{i,j}, \tag{17}$$
where $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}_s$ are respectively the covariances of the distributions of the subsets of variables in the cliques and separators. This is a direct consequence of the decomposition property for the multivariate normal distribution (see Eq. 12).

As a consequence, the ML sparse inverse covariance solution must have the form of Eq. 17 with elements given by the sample covariances, and this is indeed Eq. 6. □

Appendix C. ML solution for the Student-t distribution

Let me start from the definition of the multivariate Student-t probability density function.

Definition 4 (*Multivariate Student-t distribution*). Given of a set of random variables $\mathbf{X} \in \mathbb{R}^{p\times 1}$ the multivariate Student-t probability density function has the following canonical general expression [32]:
$$t(\mathbf{X}=\mathbf{x}) = \sqrt{\frac{1}{|\Omega|(\nu\pi)^p}}\frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^{\top}\Omega^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+p}{2}} \tag{18}$$
where $\boldsymbol{\mu} \in \mathbb{R}^{p\times 1}$ is the vector of location parameters; $\Omega \in \mathbb{R}^{p\times p}$ is a positively defined matrix called shape matrix; and $\nu > 0$ is a scalar called degrees of freedom.

The covariance matrix is defined when $\nu > 2$ and it is given by
$$\boldsymbol{\Sigma} = \frac{\nu}{\nu-2}\boldsymbol{\Omega}.$$

Assuming, $\nu > 2$, consistently with the previous notation for the normal case, I re-write the expression for the Student-t distribution in terms of the inverse covariance matrix $\mathbf{J}$.
$$t(\mathbf{X}=\mathbf{x}) = \sqrt{\frac{|\mathbf{J}|}{((\nu-2)\pi)^p}}\frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^{\top}\mathbf{J}(\mathbf{x}-\boldsymbol{\mu})}{\nu-2}\right)^{-\frac{\nu+p}{2}} \tag{19}$$

The EM construction makes use of the fact that the multivariate Student-t can be written as a normal mixture representation:
$$t(\mathbf{X}=\mathbf{x}) = \int_0^{+\infty} h(z|\frac{\nu}{2},\frac{\nu}{2})\varphi(\mathbf{x}|\boldsymbol{\mu},\frac{\nu}{\nu-2}\mathbf{J}z)dz \tag{20}$$

where

$$\varphi\left(\mathbf{x}|\boldsymbol{\mu},\frac{v}{v-2}\mathbf{J}z\right)=\sqrt{\frac{z^p v^p|\mathbf{J}|}{(2\pi(v-2))^p}}\exp\left[-\frac{z}{2}\frac{v}{v-2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{J}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{21}$$

is the multivariate normal density function with $\boldsymbol{\mu}\in\mathbb{R}^{p\times 1}$ the location parameters and z is scalar multiplying the inverse covariance matrix. Instead $h(z|\frac{v}{2},\frac{v}{2})$ is the probability density function of a gamma distribution

$$h(z|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}\ . \tag{22}$$

with both scale α and rate β parameters equal to $\frac{v}{2}$.

Let me then recap the expectation maximization (EM) approach step by step, explicitly taking into account the sparsity of $\mathbf{J}$ in our case. The EM approach proceeds into two main steps. The E-step, where and expectation function is defined; then an M-step, where it is maximized recursively.

Let me call

$$f(\mathbf{x},z|\boldsymbol{\mu},\mathbf{J},v)=h(z|\frac{v}{2},\frac{v}{2})\varphi(\mathbf{x}|\boldsymbol{\mu},\frac{v}{v-2}\mathbf{J}z). \tag{23}$$

- **E step.**
  I define the following expectation:

$$Q(\boldsymbol{\mu},\mathbf{J}|\boldsymbol{\mu}^t,\mathbf{J}^t)=\sum_{s=1}^{q}\int_0^\infty f(z|\hat{\mathbf{x}}(s),\boldsymbol{\mu}^t,\mathbf{J}^t,v)\log f(\hat{\mathbf{x}}(s),z|\boldsymbol{\mu},\mathbf{J},v)dz. \tag{24}$$

  - **M step.** I now search for the maxima of the expectation by differentiating with respect the parameters and equalling to zero.

$$\frac{\partial}{\partial\boldsymbol{\mu}}Q(\boldsymbol{\mu},\mathbf{J}|\boldsymbol{\mu}^t,\mathbf{J}^t)=\frac{v}{v-2}\sum_{s=1}^{q}\int_0^\infty zf(z|\hat{\mathbf{x}}(s),\boldsymbol{\mu}^t,\mathbf{J}^t,v)(\hat{\mathbf{x}}_s-\boldsymbol{\mu})^\top\mathbf{J}dz\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}^{t+1}}=0$$

  which, if $\mathbf{J}$ is positively defined, results in the solution

$$\boldsymbol{\mu}^{t+1}=\frac{\sum_{s=1}^{q}w_s^t\hat{\mathbf{x}}(s)}{\sum_{s=1}^{q}w_s^t}, \tag{25}$$

  with

$$w_s^t=\int_0^\infty zf(z|\hat{\mathbf{x}}(s),\boldsymbol{\mu}^t,\mathbf{J}^t,v)dz \tag{26}$$

  which can be computed explicitly. Indeed, substituting Eq. 21 and 22 one has

$$w_s^t\propto\int_0^\infty zg(z|\frac{v+p}{2},\frac{v+\frac{v}{v-2}d^2_{\hat{\mathbf{x}}(s)}}{2}dz.$$

  that is the expected value for a gamma distribution with $\alpha=\frac{v+p}{2}$ and $\beta=\frac{v+\frac{v}{v-2}d^2_{\hat{\mathbf{x}}(s)}}{2}$ which is

$$w_s^t=\frac{v+p}{v+\frac{v}{v-2}d^2_{\hat{\mathbf{x}}(s)}}. \tag{27}$$

  where the quantity

$$d^2_{\hat{\mathbf{x}}(s)}=(\hat{\mathbf{x}}(s)-\boldsymbol{\mu}^t)^\top\mathbf{J}^t(\hat{\mathbf{x}}(s)-\boldsymbol{\mu}^t), \tag{28}$$

  depends on the stage t of the EM process and therefore w_s^t must be computed recursively. Convergence is guaranteed (Theorem 2 in [28]) although it can be slow.
  In this paper the inverse covariance matrix $\mathbf{J}$ is sparse however the structure of this matrix has no relevance for the derivation of Eq. 25.
  For the derivation of $\mathbf{J}^{t+1}$ we also proceed following the same steps as for the unconstrained full case, with the only attention that the partial derivatives must be only applied over the non-zero elements with both i,j belonging to a clique:

$$\frac{\partial}{\partial J_{i,j}} Q(\boldsymbol{\mu}, \mathbf{J} | \boldsymbol{\mu}^t, \mathbf{J}^t) \tag{29}$$
$$= \frac{1}{2} \sum_{s=1}^{q} \int_0^\infty f(z|\hat{\mathbf{x}}(s), \boldsymbol{\mu}^t, \mathbf{J}^t, \nu) \left(-(\mathbf{J}^{-1})_{i,j} + z \frac{\nu}{\nu - 2} (\hat{x}_i(s) - \mu_i^t)(\hat{x}_j(s) - \mu_j^t) \right) dz \Big|_{\mathbf{J}=\mathbf{J}^{t+1}} = 0$$

resulting in the solution

$$((\mathbf{J}^{t+1})^{-1})_{i,j} = \left(\frac{\nu}{\nu - 2} \right) \frac{1}{q} \sum_{s=1}^{q} w_s^t (\hat{x}_i(s) - \mu_i^t)^\top (\hat{x}_j(s) - \mu_j^t). \tag{30}$$

In principle, I could perform the EM approach to estimate ν and, again, sparsity plays no role. However, in this paper I prefer to estimate ν from the tails of the distribution instead of using the EM approach. Then the computation is reiterated until convergence to a stable set of coefficients.

Let me now proceed with the proofs of Theorems 3 and 4 which are straightforward consequences of the previous derivation.

Proof of Theorem 3 In order to prove Theorem 3, I must demonstrate that Eqs. 8 and 9 are indeed the maximum likelihood solutions. However this is already derived in Eqs. 26 and 28, providing that the recursion procedure is convergent, but this ie guaranteed by Theorem 2 in [29]. □.

Proof of Theorem 4 Again, in order to prove Theorem 4, I must demonstrate that Eqs. 10 and 11 are the maximum likelihood solutions. However this is already derived in Eqs. 32 and 28, providing that the recursion procedure is convergent, but this ie guaranteed by Theorem 2 in [29]. □.

Appendix D. Further comparison between models

Let me here report some extra results useful for comparison between the models.

I first investigate the real data using the normal modeling instead of the Student-t. In Fig. 3 I report the log-likelihoods for real data obtained from normal models (Eq. 2) with IFN constructed using the Pearson estimate of the correlation matrix. The result for the Student-t is reported in this figure with the slashed line for comparison. We observe an overall behavior very similar to what reported for the Student-t approach (see Fig. 2), however the values of the log-likelihoods are significantly lower. This indicates that real data from financial log-returns are better modeled with Student-t multivariate probability distributions. This is not a surprise since the literature abundantly reports the inadequacy of normal modeling for financial returns, yet this result is very clean and has a referential value.

I then repeated the experiments on synthetic data generated from multivariate normal and Student-t distributions. Results for normally distributed data are reported in Fig. 4. Unsurprisingly, I observe that normal modeling with Pearson estimate of the covariance gives largest likelihoods on normal data. I also observe that the Student-t model with expectation maximization give good results similar to the normal models with Kendal estimate of the covariance. For the Student-t model with expectation maximization optimization I obtain very small differences when the Pearson's or Kendall's estimates are used. Indeed the slashed and dotted lines appear overlapping, they are however not coinciding and the Pearson's estimate give marginally better results.

Results for Student-t distributed data are reported in Fig. 5. Here, coherently, I observe that Student-t models largely outperform normal models. I also observe that contrary to the normal data case, here the Kendal's estimates of the covariances is advantageous.

In all the cases I have investigated, Glasso is outperformed by IFN-LoGo sparse models up to a certain level of sparsity and then they become equivalent.

References

[1] Andrey Nikolayevich Tikhonov, On the stability of inverse problems, Dokl. Akad. Nauk SSSR 39 (1943) 195–198.
[2] Robert Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1) (1996) 267–288.
[3] Kush R. Varshney, Karthikeyan Natesan Ramamurthy, Persistent topology of decision boundaries, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2015, pp. 3931–3935.
[4] Chao Chen, Xiuyan Ni, Qinxun Bai, Yusu Wang, A topological regularizer for classifiers via persistent homology, in: The 22nd International Conference on Artificial Intelligence and Statistics PMLR, 2019, pp. 2573–2582.
[5] Karthikeyan Natesan Ramamurthy, Kush Varshney, Krishnan Mody, Topological data analysis of decision boundaries with application to model selection, in: International Conference on Machine Learning PMLR, 2019, pp. 5351–5360.
[6] Henry Adams, Michael Moy, Topology applied to machine learning: From global to local, Front. Artif. Intell. 4 (2021) 54.
[7] Wolfram Barfuss, Guido Previde Massara, Tiziana Di Matteo, Tomaso Aste, Parsimonious modeling with information filtering networks, Phys. Rev. E 94 (6) (2016) 062306.
[8] Rosario Nunzio Mantegna, Hierarchical structure in financial markets, Eur. Phys. J. B 11 (1) (1999) 193–197, http://EconPapers.repec.org/RePEc:spr:eurphb:v:11:y:1999:i:1:p:193-197.
[9] M. Tumminello, T. Aste, T. Di Matteo, R.N. Mantegna, A tool for filtering information in complex systems, Proc. Natl. Acad. Sci. U.S.A. 102 (30) (2005) 10421–10426, https://doi.org/10.1073/pnas.0500298102.
[10] Elise A.R. Serin, Harm Nijveen, Henk W.M. Hilhorst, Wilco Ligterink, Learning from co-expression networks: possibilities and challenges, Front. Plant Sci. 7 (2016) 444.

[11] Gautier Marti, Frank Nielsen, Mikołaj Bińkowski, Philippe Donnat, A review of two decades of correlations, hierarchies, networks and clustering in financial markets, Progress in Information Geometry, 2021, pp. 245–274.
[12] T. Di Michele Tumminello, T Aste Matteo, Rosario Nunzio Mantegna, Correlation based networks of equity returns sampled at different time horizons, Eur. Phys. J. B 55 (2) (2007) 209–217.
[13] Tomaso Aste, W. Shaw, Tiziana Di Matteo, Correlation structure and dynamics in volatile markets, New J. Phys. 12 (8) (2010) 085009.
[14] Francesco Pozzi, Tiziana Di Matteo, Tomaso Aste, Spread of risk across financial markets: better to invest in the peripheries, Scientific Rep. 3 (2013) 1665.
[15] Nicoló Musmeci, Tomaso Aste, and Tiziana Di Matteo, Risk diversification: a study of persistence with a filtered correlation-network approach. arXiv preprint arXiv:1410.5621, 2014.
[16] N. Musmeci, T. Aste, T. Di Matteo, Relation between financial market structure and the real economy: Comparison between clustering methods, PLoS ONE 10 (3) (2015) e0116201.
[17] Pier Francesco Procacci, Tomaso Aste, Forecasting market states, Quantitative Finance 19 (9) (2019) 1491–1498.
[18] Alexander P. Christensen, Yoed N. Kenett, Tomaso Aste, Paul J. Silvia, Thomas R. Kwapil, Network structure of the wisconsin schizotypy scales–short forms: Examining psychometric network filtering approaches, Behavior Res. Methods 50 (6) (2018) 2531–2550.
[19] Alexander P. Christensen, Networktoolbox: Methods and measures for brain, cognitive, and psychometric network analysis in r, R J. 10 (2) (2018) 422–439.
[20] Won-Min Song, Tomaso Aste, Tiziana Di Matteo, Correlation-based biological networks, Complex Systems II, vol. 6802, International Society for Optics and Photonics, 2008, p. 680212.
[21] Won-Min Song, Tiziana Di Matteo, Tomaso Aste, Hierarchical information clustering by means of topologically embedded graphs, PloS one 7 (3) (2012) e31929.
[22] Guido Previde Massara, T. Di Matteo, Tomaso Aste, Network filtering for big data: Triangulated maximally filtered graph, J. Complex Networks 5 (2) (2017) 161, https://doi.org/10.1093/comnet/cnw015.
[23] Jerome Friedman, Trevor Hastie, Robert Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–441.
[24] Guido Previde Massara and Tomaso Aste, Learning clique forests. JMC, To be submitted, available on ArXiv abs/1905.02266, 2018.
[25] Guido Previde Massara, Building Information Filtering Networks with Topological Constraints: Algorithms and Applications (Ph.D. thesis), UCL, Computer Science, 2021.
[26] Steffen L. Lauritzen, Graphical Models, Clarendon, Oxford, 1996.
[27] Kai Wang Fang, Symmetric multivariate and related distributions, CRC Press, 2018.
[28] Mahalanobis Prasanta Chandra, et al., On the generalised distance in statistics, in: Proceedings of the National Institute of Sciences of India, vol. 2, pages 49–55, 1936.
[29] Arthur P. Dempster, Nan M. Laird, Donald B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. Roy. Stat. Soc.: Ser. B (Methodol.) 39 (1) (1977) 1–22.
[30] Christopher M. Bishop, Pattern recognition and machine learning, Springer, 2006.
[31] Cho-Jui Hsieh, Mátyás A. Sustik, Inderjit S. Dhillon, Pradeep D. Ravikumar, Quic: quadratic approximation for sparse inverse covariance estimation, J. Mach. Learn. Res. 15 (1) (2014) 2911–2947.
[32] Samuel Kotz, Saralees Nadarajah, Multivariate t-distributions and their applications, Cambridge University Press, 2004.