Undirected Graphical Model

Guoyu Zhang, Jiaqiang Li

University of Science and Technology of China

December 3, 2021

Content

Table of Contents

Motivation

- Many researches concerns to obtain the empirical evidence on variables capturing characteristics, behaviour, abilities, attitudes of people. These variables are properties of observational units.
- The objective of such studies is to improve knowledge about the association structure of the observed units via their properties, or to infer the system of variables via their association structure.

Social Network

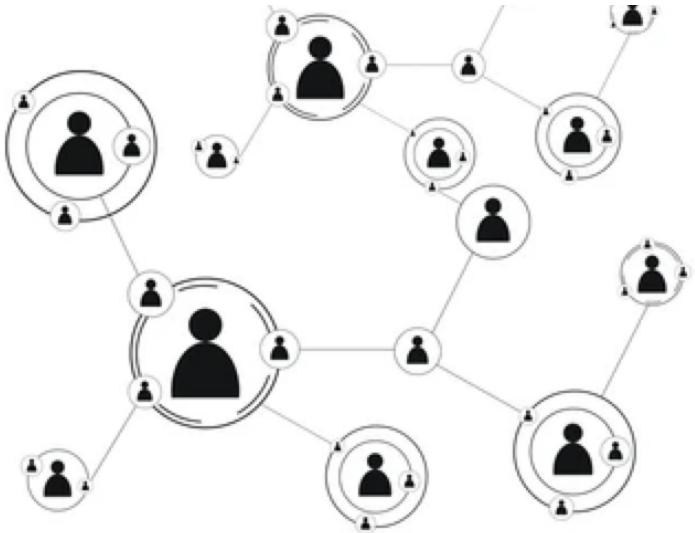


Figure: An example of social network. We want to study the relationship of people via their properties, or predict the behaviour of people via their contact network.

Motivation

- One hopes represent the structure as sparse as possible, i.e. with a few strong associations, which account for the indirect relations in the system.
- Direct and only indirect relations are key notions in studying association structures. To say that a relation between a pair of variables is indirect means that they have a substantial association given the information on other variables.
- This implies the system of variables are generally related, but only a few of these relationships are directly connected without conditioning on extra information.

Why Graphs?

There are sometimes that what we know or are interested in only concerns about the (conditional) independence among variables, rather than their accurate relationships.

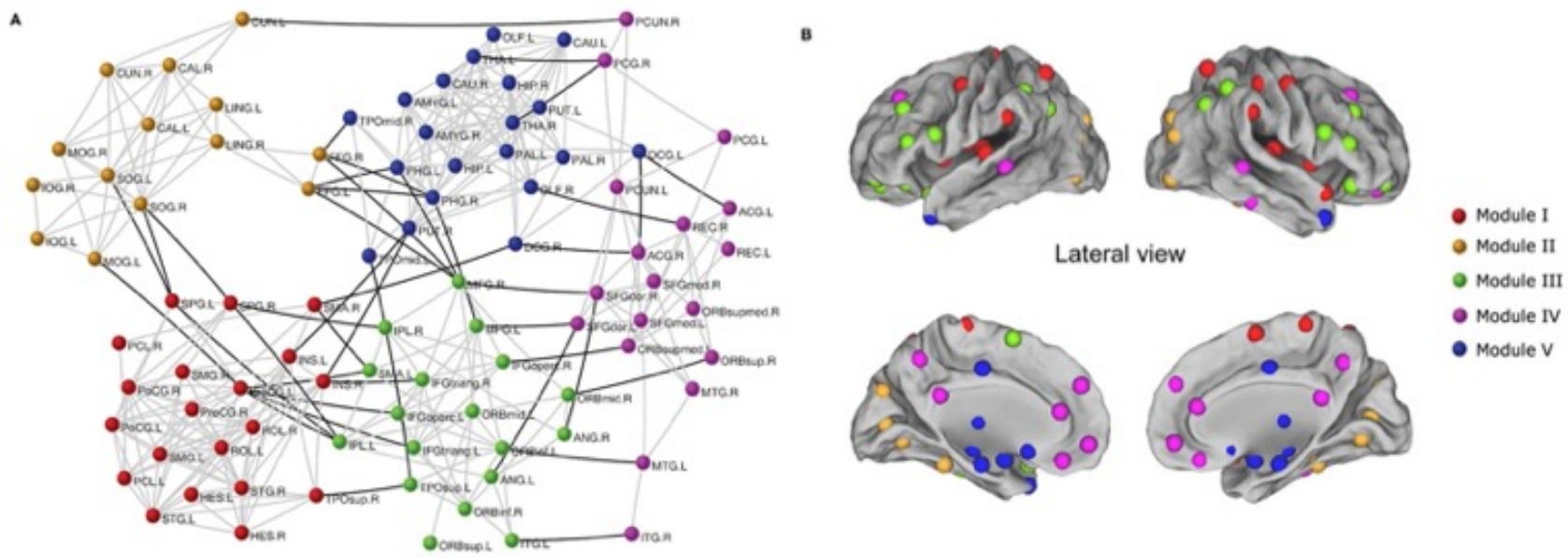


Figure: An example of graphs.

Introduction

- Graph models show our knowledge of the association structure of a given set of variables, meaning they contain the information of (conditional) independence, while still leaving enough freedom for the exact form of the distribution.
- Undirected graphical models (UGM), to be more specific, are often used in problems in which there is little causal structure to guide the construction of a directed graph but with structure of conditional independence involving.

Basic Knowledge of Graphs

- Typically, we denote an undirected graph as a tuple (V, E).
- V is the set of vertices of the graph, which we will use to represent the variables.
- E is the set of edges of the graph, which we will use to represent the 'dependence among variables' (to be defined later).

Notation: for $i, j \in V$ we use the symbol (i, j) to denote the edge from i to j whether it is in E or not.

Basic Settings

- For a d-dimensional random vector $X = (X_1, X_2, \cdots X_d)$, we need to find a graph $G = (V, E)$ to describe their probabilistic relationship.
- For each variable X_i, we use one vertex $i \in V$ to represent it.
- We hope to use whether there are edges between vertices to reflect the conditional dependence between variables given the remainders.

To achieve this, we need to the define firstly what is 'clique'.

Cliques and Maximal Cliques

Definition 1

A clique C is a subset of vertices V that are all joined by edges, i.e. for all distinct $j, k \in C$, we have $(i, j) \in E$

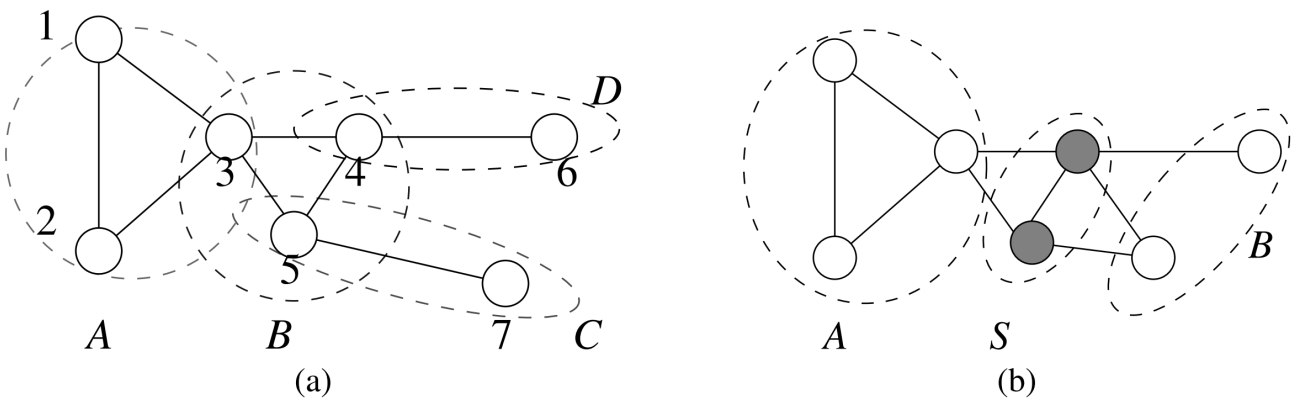


Figure: Illustration of Cliques (a).

Cliques and Maximal Cliques

Definition 2

A maximal clique C is a clique that is not a subset of any other clique.

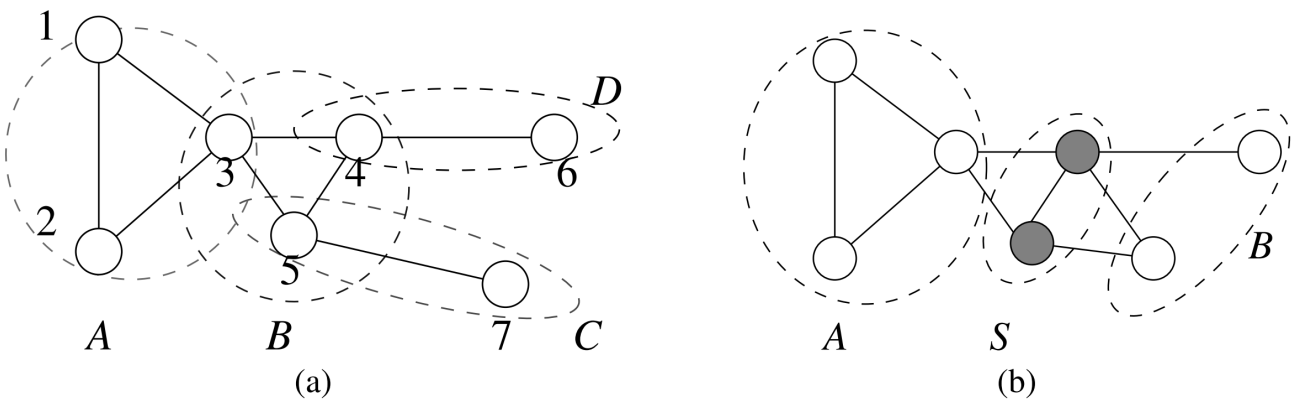


Figure: Illustration of maximal Cliques (a).

Note that any function on a clique can be seen as a function on some maximal clique.

Functions on Cliques

- We use $\mathcal{C}$ to denote the set of all cliques $C \subset V$.
- By using notation ψ_C, we mean a function on clique C, i.e.

$$\psi_C : x_C \to \mathbb{R}.$$

where $x_C := \{x_i\}_{i \in C}$. With these preparations, we are ready to give a representation for probabilistic mass via an undirected graph.

Factorization

Definition 3

We call random vector $(X_1, \cdots, X_d)$ factorizes according to graph G if we have

$$p(x_1, x_2, \cdots x_d) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where ψ_C is a non-negative function on clique C.

- Note that we can equivalently define the factorization only according to maximal cliques instead of cliques, but sometimes using the 'clique' version will bring us some convenience, as we will see in the example of Gaussian model later.

An Example

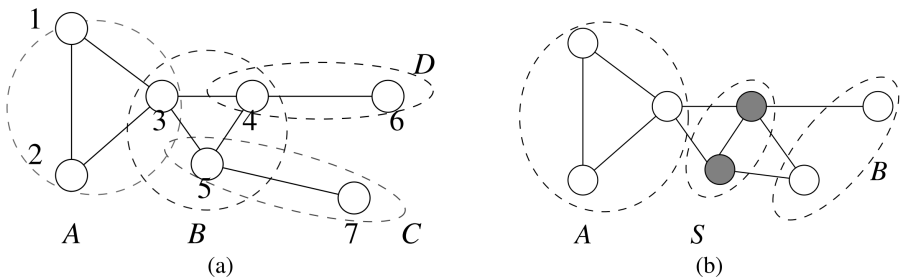


Figure: Illustration of factorization (b).

The graph to factorize the random vector does always exists (complete graph), while the density factorized by given graph (a) must have this form

$$p(x_1, \cdots x_7) \propto \psi_{123}(x_1, x_2, x_3)\psi_{345}(x_3, x_4, x_5)\psi_{46}(x_4, x_6)\psi_{57}(x_5, x_7).$$

A Question

- Till now, we have denoted the distribution via an undirected graph, but why is it that way?
- Recall our purpose: we want to define a way to easily show the conditional dependence among variables, dose this definition satisfy our request?

To illustrate this, let us first define the conditional independence 'required' by an undirected graph structures.

Vertex Cut

For a vertices subset S of V, we define the graph of removing S as $G(V \setminus S) = (V \setminus S, E(V \setminus S))$, here the set of edges are defined as

$$E(V \setminus S) = \{(j,k) \in E \mid j,k \in V \setminus S\}.$$

It is natural to expect that if the graph $G(V \setminus S)$ consists of two distinct sub-graph, they are independent (conditioned on S).

Conditional Independence induced by Undirecrted Graph

Definition 4

We say a random variable $X = (X_1, X_2, \cdots X_d)$ is Markov with respect to the undirected graph G, if $G(V \backslash S)$ is two distinct sub-graph A and B, we have $X_A \perp X_B \mid X_S$, i.e. X_A and X_B is independent conditioned on X_S.

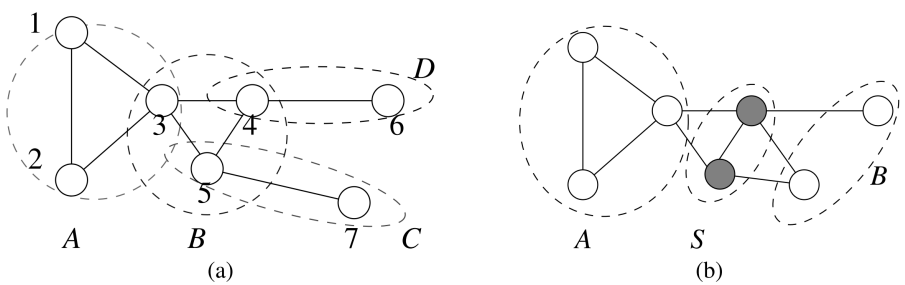


Figure: Illustration of independence induced by graph.

Markov Random Field

- A Markov random field (MRF), is the generalization of Markov chain for the random object defined on a field, which satisfies Markov properties with respect to a given graph of that field.
- The neighbourhood of node is defined via a graph, which encodes the conditional independence of the field of variables.

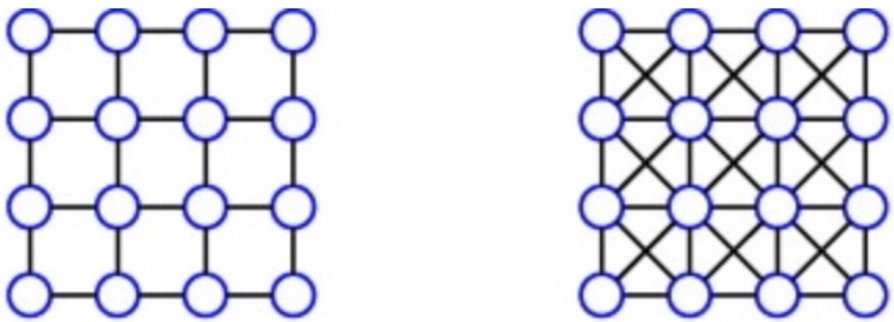


Figure: Illustration of MRF.

Markov Chains as an example

- We now consider an example: Markov chain. The states are $1, 2, \cdots k$ and we can easily draw a graph that contains these k nodes, with edges only existing between $(i, i+1)$
- By the graph, we might write the probability as

$$p(x_1, \cdots x_k) \propto p(x_1)p(x_2|x_1)\cdots p(x_k|x_{k-1})$$

  which means it factorize according to graph
- Now if we cut the nodes from some i to j, the remaining two parts are independent conditioned on the variables i to j, which means this it is also Markov with the graph.

Is this just a coincidence that these two properties are satisfied simultaneously? The famous Hammersley–Clifford theorem[6] tells you 'NO'!

Hammersley–Clifford equivalence

Theorem 1

For a given undirected graph G and random variables $X = (X_1, X_2, \cdots X_d)$, with strictly positive density p, the following two properties are equivalent:
(1) X factorize as the graph G.
(2) X is Markov with the graph G.

This interesting theorem tells us two things:

- Graph models can represent the conditional dependence relationships among variables.
- The only way to define the probability induced by a graph, if request to keep independence, is the way we used above.

Simple proof for necessity: A closer look at Markov chains

To give the idea of the proof for the necessity, i.e. (2) contains (1), we look closer at the simplest situation of 'Markov chain':

- Consider a 3-state Markov chain, with graph

$$G = (\{1, 2, 3\}, \{(1, 2), (2, 3)\}).$$

- According to the factorization, we can write

$$p(x_1, x_2, x_3) \propto \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)$$

-

$$p(x_2) \propto \sum_{x_1, x_3} \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)$$
$$= \sum_{x_1} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{23}(x_2, x_3)$$

.

Simple proof for necessity: A closer look at Markov chains

- $$p(x_1, x_{x3}|x_2) \propto \frac{\psi_{12}(x_1, x_2)}{\sum_{x_1} \psi_{12}(x_1, x_2)} \cdot \frac{\psi_{23}(x_2, x_3)}{\sum_{x_3} \psi_{23}(x_2, x_3)}$$
$$= p(x_1|x_2)p(x_3|x_2)$$

and the conditional independence is obtained.

Gaussian Graph Model

We say a d-dimension random vector X obeys multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, if X has the density

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1}(x-\mu)}$$

From now on, let $\mu = 0$.
We use the notation σ_{ij} to denote the (i, j) element of Σ and σ^{ij} for the (i, j) element of Σ^{-1}, then we can rewrite the density function as:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}\sum_{i,j} \sigma^{ij} x_i x_j}.$$

Conditional Independence for multivariate normal distribution

- By using the notations we define above, let us see what will happen if $\sigma^{12} = 0$.
- Under this condition, we have

$$f(x) \propto e^{-\frac{1}{2}x_1(\sum_{j\neq 1,2} 2\sigma^{1j}x_j + \sigma^{11}x_1)} e^{-\frac{1}{2}x_2(\sum_{i\neq 1,2} 2\sigma^{2i}x_i + \sigma^{22}x_2)} e^{-\frac{1}{2}\sum_{i,j\neq 1,2}\sigma^{ij}x_i x_j},$$

  which exactly tells us that $X_1 \perp X_2 | X_3, \cdots X_d$.
- Similarly, we can prove that $\sigma^{ij} = 0$ if and only if that $X_i \perp X_j | X_{-i,-j}$, here $X_{-i,-j}$ means all the other components except i, j. Even better, similar conclusion holds for sub-matrix of Σ^{-1}.

Gaussian Graph Model

With above preparations, we can now define the structure of Gaussian graph models:

- For $X \sim \mathcal{N}(0, \Sigma)$, we concnnect an undirected graph model $G = (V, E)$ to it
- Vertex i in V represent the i-th component of X.
- There is an edge between i and j if and only if $\sigma^{ij} \neq 0$.

Then we can easily show that with these settings, X factorizes with the graph G, thus is also Markov with G.
Note that if we require the factor is defined as functions of maximal cliques, it will be much more inconvenient to represent this.

Covariance Estimation with Gaussian Graph Model

- A natural question arises: when knowing the structure of Gaussian graph model and samples $X_1, \cdots X_n$, how can we estimate the covariance matrix now?
- According to our construction, knowing the structure of Gaussian graph model is equivalent to knowing which $\sigma^{ij} = 0$ if $(i, j) \notin E$.
- At this time, the usual unbiased estimation

$$S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})' (X_i - \bar{X})$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_n$, might not be our ideal option since the (i, j)th element of $\hat{\Theta} := S^{-1}$ may not be 0 if $(i, j) \notin E$.

Covariance Estimation with Gaussian Graph Model

Let $J = \{(i,j) \mid \sigma^{ij} = 0\}$, the set of all the conditional independent pairs, and $I = J^c$, the remaining set of J, then one seemingly crazy way to construct our new estimator $\tilde{S}$ is as follows:

- $W|_I = S|_I$, i.e. W is the same as S at the position of I.
- The (i,j) element of the inverse of W is 0 for any $(i,j) \in J$.

Clearly, we might ask these two questions:

- Is there really existing such an estimator that satisfying the above conditions? If yes, will it be unique?
- Why we choose our estimator like this? What properties it might own? [1]

Covariance Estimation with Gaussian Graph Model

Theorem 2

Given positive definite matrices L and M defined on the vertices V of a graph G = (V,E) there exists a unique positive definite matrix K such that
(1) $K_{ij} = L_{ij}$ *if* $(i,j) \in E$ *or i=j,*
(2)$(K^{-1})_{ij} = M_{ij}$ *if* $(i,j) \notin E$ *and* $i \neq j$.

If there exists one positive definite matrix that agrees with S at all the positions $(i,j) \in I$, there exists one and only one positive definite matrix satisfies additionally that the inverse of which is 0 in position J.

Discussion on the above theorem

- Clearly, the above theorem can actually be generalized by not assuming zero on position J.
- An interesting way to view this theorem is given the marginal distribution of all maximal cliques, is there a joint distribution of all components owning these marginals and graph structure?
- The existence can be proven in a general exponential families setting, but instead, we will give the algorithm of how to find it to prove the existence.

The Maximum of Likelihood Property

With almost the same observation, we can show the following maximum likelihood property:

Theorem 3

Among all the multivariate normal distributions whose inverse covariance matrix are 0 at J, our choice of W is the MLE of covariance matrix.

By writing down the likelihood, we can easily find that the elements of S of position I is sufficient statistics, and this theorem just tells us how to use this sufficient statistics to construct MLE.

The Maximum of Likelihood Property

Proof.

By Lagrange multiplier, the log-likelihood of the data can be written as

$$l_C(\Theta) = \log\det\Theta - \text{trace}(S\Theta) - \sum_{(i,j)\in J} \gamma_{ij}\theta_{jk}.$$

The gradient equation is

$$\Theta^{-1} - S - \Gamma = 0 \tag{1}$$

Γ is a matrix of Lagrange parameters with nonzero values for all pairs not in E. Therefore, $W = S + \Gamma$, satisfying that $W|_I = S|_I$. □

The general cyclic algorithm

With all these good properties, a natural question might come: how can we really construct W from S?
Fortunately, there is actually an algorithm that dealing with such kind of problem:

- Given two positive definite matrices P,Q of order d, and two disjoint set I,J such that $I \bigcup J = \{(i,j) \mid 1 \leq i \leq j \leq d\}$
- How can we find a positive definite matrix F so that

$$F|_I = P|_I, \quad F^{-1}|_J = Q^{-1}|_J$$

The algorithm is called general cyclic algorithm [5].

General idea for general cyclic algorithm

We now give the main idea for such a powerful algorithm:

- First we need to find some sets $I_1, I_2, \cdots I_m$ (not necessarily distinct) so that their union is just I.
- Let $F_0 = Q^{-1}$, we try to construct a sequence of F_n so that

$$F_n|_{I_k} = P|_{I_k}, \quad F_n^{-1}|_{I_k^c} = F_{n-1}^{-1}|_{I_k^c}$$

  where $n = k(mod\ m)$.
- While maintaining the second condition, we cyclic through I_m so that the first condition is also satisfied eventually.

When applying this to our problem, we either fix the inverse of matrix to be 0 at J and forcing the I position to become S, or firstly fixing the position of I to be the same as S and making the inverse matrix to be 0 at J.

Ising Model

Consider a vector $X = (X_1, \ldots, X_d)$ of binary random variables, with each $X_j \in \{0, 1\}$. Given an undirected graph $G = (V, E)$, it posits a factorization of the form

$$p(x_1, \ldots, x_d; \theta^*) = \frac{1}{Z(\theta^*)} \exp\left\{ \sum_{j \in V} \theta^*_j x_j + \sum_{(j,k) \in E} \theta^*_{jk} x_j x_k \right\}$$

where the parameter θ^*_j is associated with vertex $j \in V$, and the parameter θ^*_{jk} is associated with edge $(j, k) \in E$ and $x_j \in \{-1, 1\}$.

Ising Model

The quantity $Z(\theta^*)$ is a constant that serves to enforce that the probability mass function p normalizes properly to one; more precisely, we have

$$Z(\theta^*) = \sum_{x \in \{0,1\}^d} \exp\left\{ \sum_{j \in V} \theta^*_j x_j + \sum_{(j,k) \in E} \theta^*_{jk} x_j x_k \right\}.$$

In general, the parameters θ^* would not possess nice properties as Gaussian cases, which requites gradient-based optimization, Monte Carlo sampling or variational algorithm for statistical inference.

Table of Contents

Gaussian Graphical Model

Undirected Graphical Model (UGM) characterizes the relation between variables. For a graph (V, E), $\{i, j\} \notin E$ means then variables **conditionally independent** (CI) given the other variables.
Let $\vec{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the **precision matrix** is defined as $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$.
If the ijth component of $\boldsymbol{\Theta}$ is zero, then variables i and j are CI.

- If we have known the structure of UGM related to a Gaussian distribution, can we estimate $\boldsymbol{\Sigma}$ better?
- Now we focus on extracting the MLE of covariance ($\boldsymbol{W}$), given the graph constraint (V, E).
- Specifically, we can make $\boldsymbol{W}_{ij} = \boldsymbol{S}_{ij}$ if $\{i, j\} \in E$ or $i = j$, $\boldsymbol{W}_{ij}^{-1} = 0$ if $\{i, j\} \notin E$, where $\boldsymbol{S}$ represent the empirical covariance matrix.

Multiple Liner Regression

Partition $\vec{X} = (\vec{Z}, Y)$ where $\vec{Z} = (X_1, \ldots, X_{p-1})$, $Y = X_p$. Partition $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{ZZ} & \boldsymbol{\sigma}_{ZY} \\ \boldsymbol{\sigma}_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

For simplicity, we denote $\boldsymbol{\Sigma}_{ZZ}$ as $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\sigma}_{ZY}$ as $\boldsymbol{\sigma}_{12}$, similar for other elements in $(X_1, \cdots, X_p)$.
Partition $\boldsymbol{\Theta}$ in the same way, since $\boldsymbol{\Sigma}\boldsymbol{\Theta} = \boldsymbol{I}$, we obtain

$$\boldsymbol{\theta}_{12} = -\theta_{22}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_{12} = -\theta_{22}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} := \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_{12}$ is exactly the regression coefficient of Y onto $\vec{Z}$. Hence $\boldsymbol{\beta} = -\boldsymbol{\theta}_{12}/\theta_{22}$.

Multiple Liner Regression

Now we focus on iteratively updating the columns of $\boldsymbol{W}$. Let

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}^T & w_{22} \end{pmatrix}.$$

Let $\hat{\boldsymbol{\Theta}}$ be the estimation of $\boldsymbol{\Theta}$. Since $\boldsymbol{W}\hat{\boldsymbol{\Theta}} = \boldsymbol{I}$,

$$\boldsymbol{w}_{12} = -\boldsymbol{W}_{11}\hat{\boldsymbol{\theta}}_{12}/\hat{\theta}_{22} := \boldsymbol{W}_{11}\hat{\boldsymbol{\beta}}.$$

Focus on the upper right block of equation (1), we have

$$\boldsymbol{w}_{12} - \boldsymbol{s}_{12} - \boldsymbol{\gamma}_{12} = \boldsymbol{0}.$$

By given $\boldsymbol{W}_{11}$, we can solve $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{W}_{11}\hat{\boldsymbol{\beta}} - \boldsymbol{s}_{12} - \boldsymbol{\gamma}_{12} = \boldsymbol{0}$.

Modified Regression

Remove the rows and columns that the corresponding predictors don't connect to the respond, we have

$$\boldsymbol{W}_{11}^{*}\boldsymbol{\beta}^{*} - \boldsymbol{s}_{12}^{*} = \mathbf{0}, \tag{2}$$

then $\boldsymbol{\beta}^{*} = {\boldsymbol{W}_{11}^{*}}^{-1}\boldsymbol{s}_{12}^{*}$. Pad the zeros in corresponding locations to give $\hat{\boldsymbol{\beta}}$.

A Modified Regression Algorithm

Algorithm 17.1 *A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.*

1. Initialize $\mathbf{W} = \mathbf{S}$.
2. Repeat for $j = 1, 2, \ldots, p, 1, \ldots$ until convergence:
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.
   (b) Solve $\mathbf{W}_{11}^{*}\beta^{*} - s_{12}^{*} = 0$ for the unconstrained edge parameters β^{*}, using the reduced system of equations as in (17.19). Obtain $\hat{\beta}$ by padding $\hat{\beta}^{*}$ with zeros in the appropriate positions.
   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$
3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = s_{22} - w_{12}^{T}\hat{\beta}$.

Why this algorithm converges to our target?

Algorithm 17.1 is a special case of the general cyclic algorithm [5].

- $\boldsymbol{W}^n$ is the matrix after n iterations and $\boldsymbol{W}^0 = \boldsymbol{S}$.
- Suppose $\boldsymbol{W}^n$ is invertible and positive definite, by induction we know $\boldsymbol{W}^{n+1}$ is invertible and positive definite.
- Let $\boldsymbol{\Theta}^n = (\boldsymbol{W}^n)^{-1}$, then $\{\boldsymbol{\Theta}^n\}$ satisfies that:

$$\boldsymbol{W}^n_{ij} = \boldsymbol{W}^{n-1}_{ij} = \boldsymbol{S}_{ij}, \quad (i,j) \notin \tilde{E}_{n'}, \tag{3}$$

$$\boldsymbol{\Theta}^n_{ij} = \boldsymbol{\Theta}_{ij} = 0, \quad (i,j) \in \tilde{E}_{n'}. \tag{4}$$

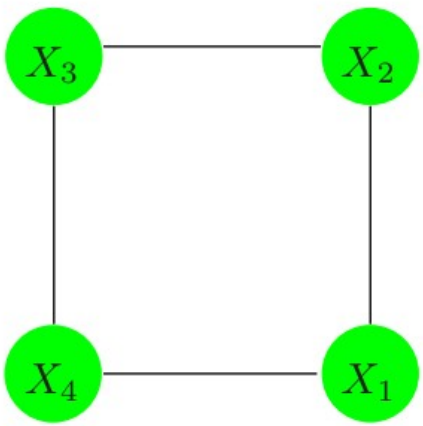
where $\tilde{E}_k = \{(k,j) | j \in V, (k,j) \notin E\}, k = 1, \ldots, p$, is the edge set without connecting to node k. $n' = n(mod\ p)$.

Iteration Process

- Suppose (3),(4) hold after $(n-1)$th iterations, we focus on the n-th iteration. Noting that $\boldsymbol{W}_{11}$ would not change in nth iterations. We just need to focus on $\boldsymbol{w}_{12}$.
- Let A be the node set connecting target, and B be the node set without linking target, and

$$(\boldsymbol{w}_{12})_A = (\boldsymbol{W}_{AA}, \boldsymbol{W}_{AB}) \times \begin{pmatrix} \boldsymbol{W}_{AA}^{-1} \boldsymbol{s}_{12}^* \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{s}_{12}^*.$$

A Simple Example



$$\mathbf{S} = \begin{pmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{pmatrix}$$

$$\hat{\mathbf{\Sigma}} = \begin{pmatrix} 10.00 & 1.00 & 1.31 & 4.00 \\ 1.00 & 10.00 & 2.00 & 0.87 \\ 1.31 & 2.00 & 10.00 & 3.00 \\ 4.00 & 0.87 & 3.00 & 10.00 \end{pmatrix}, \quad \hat{\mathbf{\Sigma}}^{-1} = \begin{pmatrix} 0.12 & -0.01 & 0.00 & -0.05 \\ -0.01 & 0.11 & -0.02 & 0.00 \\ 0.00 & -0.02 & 0.11 & -0.03 \\ -0.05 & 0.00 & -0.03 & 0.13 \end{pmatrix}$$

Estimating Graphs from Data

- Meinshausen et al. (2006) apply lasso to estimate which components of $\mathbf{\Theta}$ are nonzero.
- Fit a lasso regression using each variable as the response and the others as predictors.
- Include an edge $\{i, j\}$ in the graph if the estimated precision matrix of variable i on j is nonzero.

Graphical Lasso

- Consider maximizing the penalized log-likelihood

$$\max_{\boldsymbol{\Theta}>0}\{\log\det\boldsymbol{\Theta} - \mathit{trace}(\boldsymbol{S}\boldsymbol{\Theta}) - \lambda\|\boldsymbol{\Theta}\|_1\}, \tag{5}$$

- The gradient equation (stationary condition) is

$$\boldsymbol{\Theta}^{-1} - \boldsymbol{S} - \lambda\boldsymbol{\Gamma} = \boldsymbol{0},$$

  where $\boldsymbol{\Gamma} = \mathit{Sign}(\boldsymbol{\Theta})$ with $\mathrm{Sign}(\theta_{jk})$=sign$(\theta_{jk})$ if $\theta_{jk} \neq 0$, else $\mathrm{Sign}(\theta_{jk}) \in [-1, 1]$, which is the sub-gradient of $\|\boldsymbol{\Theta}\|_1$.
- It's easy to show that $\boldsymbol{w}_{22} = \boldsymbol{s}_{22} + \lambda$.

Graphical Lasso

- Similar with the previous section

$$\boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} + \lambda \text{Sign}(\boldsymbol{\beta}) = \boldsymbol{0} \tag{6}$$

- Note that

$$\frac{\partial}{\partial \beta}(\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{s}_{12} + \lambda\|\boldsymbol{\beta}\|_1) = \boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{s}_{12} + \lambda \text{Sign}(\boldsymbol{\beta}).$$

The system above is exactly equivalent to a lasso regression:

$$\min_{\boldsymbol{\beta}} \{\frac{1}{2}\|\boldsymbol{V}^{1/2}\boldsymbol{\beta} - \boldsymbol{V}^{-1/2}\boldsymbol{s}_{12}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\}, \tag{7}$$

where $\boldsymbol{V} = \boldsymbol{W}_{11}$.

Graphical Lasso

Friedman et al. [2] use the pathwise coordinate descent method to solve the modified lasso problem at each stage.

- Focus on the jth row of (6) and fix $\hat{\beta}_k, (k \neq j)$.

$$\boldsymbol{V}_{jj}\hat{\beta}_j + \sum_{k \neq j} \boldsymbol{V}_{jk}\hat{\beta}_k - (\boldsymbol{s}_{12})_j + \lambda Sign(\hat{\beta}_j) = 0$$

$$\hat{\beta}_j = \lambda[((\boldsymbol{s}_{12})_j - \sum_{k \neq j} \boldsymbol{V}_{jk}\hat{\beta}_k)/\lambda - Sign(\hat{\beta}_j)]/\boldsymbol{V}_{jj}$$

- Since $Sign(t) \in [-1, 1]$ if t=0, we obtain

$$\hat{\beta}_j = S\left((\boldsymbol{s}_{12})_j - \sum_{k \neq j} \boldsymbol{V}_{jk}\hat{\beta}_k, \lambda\right)/\boldsymbol{V}_{jj},$$

where S is the soft-threshold operator:

$$S(x, \lambda) = sign(x)(|x| - \lambda)_+.$$

Graphical Lasso

Algorithm 17.2 *Graphical Lasso.*

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.

2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:

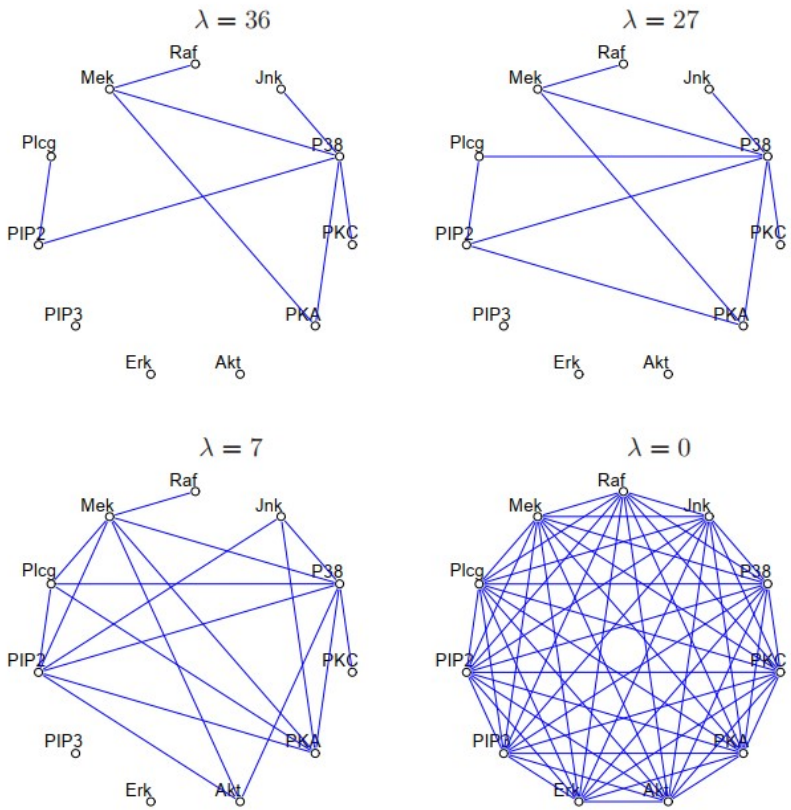
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.

   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.

   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$

3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

Model Selection



Model Selection

The graph becomes more sparse as the penalty parameter is increased.
We can use BIC for model selection.For a given penalty parameter λ,

$$BIC(\lambda) = -\log\det(\hat{\mathbf{\Theta}}(\lambda)) + trace(\hat{\mathbf{\Theta}}\boldsymbol{S}) + \frac{\log n}{n}\sum_{i\leq j}\hat{e}_{ij}(\lambda),$$

where $\hat{e}_{ij} = 0$ if $\hat{\theta}_{ij} = 0$, and $\hat{e}_{ij} = 1$ otherwise [7].

The dual problem of Graphical Lasso

Lemma 1

The primal problem (5) and its stationarity conditions are equivalent to the stationary conditions for the box-constrained SDP (semi-definite programming) [4].

$$\max_{\|\tilde{\boldsymbol{\Gamma}}\|_\infty \leq \lambda} \log\det(\boldsymbol{S} + \tilde{\boldsymbol{\Gamma}}) + p, \tag{8}$$

under the transformation $\boldsymbol{S} + \tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Theta}^{-1}$.

- Rewrite gradient conditions of (5): $-(\boldsymbol{S} + \lambda\boldsymbol{\Gamma})^{-1} + \boldsymbol{\Theta} = \boldsymbol{0}$. Let $\tilde{\boldsymbol{\Gamma}} = \lambda\boldsymbol{\Gamma}$ then $\|\tilde{\boldsymbol{\Gamma}}\|_\infty \leq \lambda$. Let $\boldsymbol{P} = abs(\boldsymbol{\Theta})$.

- Such $\tilde{\mathbf{\Gamma}}, \boldsymbol{P}$ satisfy the following equations:

$$\begin{aligned} -(\boldsymbol{S}+\tilde{\mathbf{\Gamma}})^{-1}+\boldsymbol{P}*sign(\tilde{\mathbf{\Gamma}})&=\mathbf{0} \\ \boldsymbol{P}*(abs(\tilde{\mathbf{\Gamma}})-\lambda \mathbf{1}_p\mathbf{1}_{p'})&=\mathbf{0} \\ \|\tilde{\mathbf{\Gamma}}\|_\infty &\leq \lambda, \end{aligned} \tag{9}$$

  where $\boldsymbol{P}$ is a symmetric $p \times p$ matrix with non-negative entries and the operator '*' denotes element-wise product.
  In fact, (9) are the KKT conditions for the box-constrained SDP (8).
- Take $\mathbf{\Theta}=\boldsymbol{P}*sign(\tilde{\mathbf{\Gamma}}), \mathbf{\Gamma}=\tilde{\mathbf{\Gamma}}/\lambda$, conditions (9) imply $-(\boldsymbol{S}+\lambda\mathbf{\Gamma})^{-1}+\mathbf{\Theta}=\mathbf{0}$.

Lemma 2

Assume $\boldsymbol{W}_{11} > 0$. *The stationarity equations*

$$-\boldsymbol{W}_{11}\hat{\boldsymbol{\beta}} + \boldsymbol{s}_{12} + \lambda\hat{\boldsymbol{\gamma}}_{12} = \boldsymbol{0}, \tag{10}$$

correspond to the solution of

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{W}_{11}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{s}_{12} + \lambda\|\boldsymbol{\beta}\|_1. \tag{11}$$

Solving it is equivalent to solving the following box-constrain:

$$\min_{\tilde{\boldsymbol{\gamma}}\in\mathbb{R}^{p-1},\|\tilde{\boldsymbol{\gamma}}\|_\infty\leq\lambda} \frac{1}{2}(\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}})'\boldsymbol{W}_{11}^{-1}(\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}}), \tag{12}$$

where $\hat{\boldsymbol{\beta}}$ *and* $\tilde{\boldsymbol{\gamma}}_{12}$ *are related by*

$$\hat{\boldsymbol{\beta}} = \boldsymbol{W}_{11}^{-1}(\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}}_{12}). \tag{13}$$

Proof of Lemma 2

- It's easy to find (10) is the KKT condition for problem (11) since $\boldsymbol{\beta} = -\boldsymbol{\theta}_{12}/\theta_{22}, Sign(\boldsymbol{\beta}) = -Sign(\boldsymbol{\theta}_{12})$.
- Write (10) as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{W}_{11}^{-1}(\boldsymbol{s}_{12} + \lambda\hat{\gamma}_{12}) = \mathbf{0} \tag{14}$$

Suppose $\hat{\boldsymbol{\beta}}, \hat{\gamma}_{12}$ satisfy (14), then $\tilde{\gamma}_{12} := \lambda\hat{\gamma}_{12}$, $\tilde{\boldsymbol{p}}_{12} := abs(\hat{\boldsymbol{\beta}})$ satisfy

$$\begin{aligned} \boldsymbol{W}_{11}^{-1}(\boldsymbol{s}_{12} + \tilde{\gamma}_{12}) + \tilde{\boldsymbol{p}}_{12} * sign(\tilde{\gamma}_{12}) &= \mathbf{0} \\ \tilde{\boldsymbol{p}}_{12} * (abs(\tilde{\gamma}_{12}) - \lambda\mathbf{1}_{p-1}) &= \mathbf{0} \\ \|\tilde{\gamma}_{12}\|_\infty &\leq \lambda, \end{aligned} \tag{15}$$

which are KKT conditions of (12).

- If $\tilde{\gamma}_{12}, \tilde{\boldsymbol{p}}_{12}$ satisfy (15), then

$$\hat{\gamma}_{12} = \tilde{\gamma}_{12}/\lambda, \quad \hat{\boldsymbol{\beta}} = -\tilde{\boldsymbol{p}}_{12} * sign(\hat{\gamma}_{12})$$

satisfy (14). □

Moreover, consider solving (8) for the block $\tilde{\gamma}_{12}$ holding the rest of $\tilde{\boldsymbol{\Gamma}}$ fixed. From (9), we have

$$\begin{aligned} -\boldsymbol{\theta}_{12} + \boldsymbol{p}_{12} * sign(\tilde{\gamma}_{12}) &= \mathbf{0} \\ \boldsymbol{p}_{12} * (abs(\tilde{\gamma}_{12}) - \lambda \mathbf{1}_{p-1}) &= \mathbf{0} \\ \|\tilde{\gamma}_{12}\|_\infty &\leq \lambda. \end{aligned} \tag{16}$$

Since $\hat{\boldsymbol{\beta}} = -\boldsymbol{\theta}_{12}/\theta_{22}, \theta_{22} > 0$, let $\tilde{\boldsymbol{p}}_{12} = \boldsymbol{p}_{12}/\theta_{22}$, (16) are equivalent with conditions (15). By lemma 2, GLASSO solves (15), hence (16). So GLASSO performs dual block updates for the dual of the graphical problem.

DP-GLASSO

- Instead of updating $\boldsymbol{W}$, now we update $\boldsymbol{\Theta}$ directly. Similar as before, we obtain

$$\boldsymbol{\Theta}_{11}^{-1}\boldsymbol{\theta}_{12}w_{22} + \boldsymbol{s}_{12} + \lambda\boldsymbol{\gamma}_{12} = \boldsymbol{0}, \tag{17}$$

  where $w_{22} = s_{22} + \lambda$ is fixed.
- With $\boldsymbol{\alpha} = \boldsymbol{\theta}_{12}w_{22}, \boldsymbol{\Theta}_{11} > 0$, By Lemma 2, it's equivalent to solve:

$$\min_{\tilde{\boldsymbol{\gamma}}\in\mathbb{R}^{p-1}, \|\tilde{\boldsymbol{\gamma}}\|_\infty \leq \lambda} \frac{1}{2}(\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}})'\boldsymbol{\Theta}_{11}(\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}}) \tag{18}$$

  The optimal solutions of (17) and (18) are related by

$$\hat{\boldsymbol{\alpha}} = -\boldsymbol{\Theta}_{11}(\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}}).$$

DP-GLASSO

- We update $\hat{\boldsymbol{\theta}}_{12}$ as $\hat{\boldsymbol{\alpha}}/w_{22}$.
- θ_{22} is updated by

$$\hat{\theta}_{22} = \frac{1 - (\boldsymbol{s}_{12} + \tilde{\boldsymbol{\gamma}})' \hat{\boldsymbol{\theta}}_{12}}{w_{22}}. \tag{19}$$

DP-GLASSO

Algorithm 3 DP-GLASSO algorithm

1. Initialize $\boldsymbol{\Theta} = \text{diag}(\mathbf{S} + \lambda \mathbf{I})^{-1}$.
2. Cycle around the columns repeatedly, performing the following steps till convergence:
   (a) Rearrange the rows/columns so that the target column is last (implicitly).
   (b) Solve (18) for $\tilde{\boldsymbol{\gamma}}$ and update
   $$\hat{\boldsymbol{\theta}}_{12} = -\boldsymbol{\Theta}_{11}(\mathbf{s}_{12} + \tilde{\boldsymbol{\gamma}})/w_{22}$$
   (c) Solve for θ_{22} using (19).
   (d) Update the working covariance $\mathbf{w}_{12} = \mathbf{s}_{12} + \tilde{\boldsymbol{\gamma}}$.

Nonparanormal Case

- Han Liu et al.(2012) propose *the nonparanormal SKEPTIC* to estimate high-dimensional UGM efficiently and robustly [3].
- Their simulation suggests that the nonparanormal SKEPTIC can be used as a safe replaced for Gaussian based estimator, even when the data are truly Gaussian.
- Idea: Use rank-based correlation coefficient estimators (Spearman's rho, Kendall's tau) which are invariant under monotone transformations.

Nonparanormal

Let $\vec{f} = \{f_1, \cdots, f_d\}$ be a set of monotone univariate functions and let $\mathbf{\Sigma}^0 \in \mathbb{R}^{p \times p}$ be a positive-definite correlation matrix with $\text{diag}(\mathbf{\Sigma}^0) = \mathbf{1}$.

Definition 5 (nonparanormal)

We say a p-dimensional random variable $\vec{X} = (X_1, \cdots, X_p)^T$ *has a* ***nonparanormal*** *distribution* $\vec{X} \sim NPN_p(\vec{f}, \mathbf{\Sigma}^0)$ *if*

$$\vec{f}(\vec{X}) := (f_1(X_1), \cdots, f_p(X_p))^T \sim N_p(\mathbf{0}, \mathbf{\Sigma}^0).$$

- Monotone transformation won't change conditional independence between variables, the graph structure is invariant under such transformations.

Estimating marginal transformations

Assume $x^1, \ldots, x^n$ are realizations of $\vec{X} \sim NPN_p(\vec{f}, \mathbf{\Sigma}^0)$, here $x_j^i = (x_1^i, \ldots, x_p^i)$. Estimating marginal transformations is useful for calculating the likelihood of a nonparnormal fit.
Let $\tilde{F}_j(t) = \frac{1}{n} \sum_{i=1}^n I(x_j^i \leq t)$ be the emperical distribution of X_j.
Estimate the marginal transformation f_j by

$$\tilde{f}_j(x) := \Phi^{-1}\big(T_{\delta_n}[\tilde{F}_j(x)]\big),$$

where Φ is the distribution function of $N(0,1)$ and the truncation operator

$$T_{\delta_n} := \delta_n \cdot I(x < \delta_n) + x \cdot I(\delta_n \leq x \leq 1-\delta_n) + (1-\delta_n) \cdot I(x > 1-\delta_n),$$

is used to alleviate the instability of $\Phi^{-1}(t)$ for too large or small t.

Normal-score rank correlation

- The normal-score rank correlation coefficient of $\tilde{f}_j(x)$ could be used directly in glasso. However, it appears that their efficiency cannot be generalized to the high-dimensional setting.
- The reason is that the standard Gaussian quantile function diverges very quickly when it is evaluated at a point close to 1.

Rank-based Correlation Coefficient Estimators

Let r_j^i be the rank of x_j^i among $x_j^1, \cdots, x_j^n$, $\bar{r}_j = \frac{1}{n}\sum_{i=1}^n r_j^i = \frac{n+1}{2}$.

- Spearman's rho

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_j^i - \bar{r}_j)(r_k^i - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_j^i - \bar{r}_j)^2 \cdot \sum_{i=1}^n (r_k^i - \bar{r}_k)^2}}$$

- Kendall's tau

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} sign\big((x_j^i - x_j^{i'})(x_k^i - x_k^{i'})\big).$$

- The two statistics are the same for $\vec{X}$ and $\vec{f}(\vec{X})$.

SKEPTIC

Lemma 3 (Kendall (1948), Kruskal (1958))

Assume $\vec{X} \sim NPN_p(\vec{f}, \mathbf{\Sigma}^0)$, *we have* $\Sigma^0_{jk} = 2\sin\left(\frac{\pi}{6}\rho_{jk}\right) = \sin\left(\frac{\pi}{2}\tau_{jk}\right)$.

Motivated by this lemma, define $\hat{\boldsymbol{S}}^{\rho} = [\hat{S}^{\rho}_{jk}]$ for the unknown correlation matrix $\mathbf{\Sigma}^0$ as follows:

$$\hat{S}^{\rho}_{jk} = \begin{cases} 2\sin\left(\frac{\pi}{6}\hat{\rho}_{jk}\right), & j \neq k, \\ 1, & j = k. \end{cases}$$

Similarly we can define $\hat{\boldsymbol{S}}^{\tau} = [\hat{S}^{\tau}_{jk}]$.

- Estimate $\hat{\boldsymbol{S}}$ by SKEPTIC instead of calculating Pearson correlation.

SKEPTIC with Graphical Lasso

Plug the estimated correlation matrix $\hat{\boldsymbol{S}}$ into the graphical lasso:

$$\hat{\boldsymbol{\Theta}}^{glasso} = \arg\min_{\boldsymbol{\Theta}>0} \{trace(\hat{\boldsymbol{S}}\boldsymbol{\Theta}) - \log\det\boldsymbol{\Theta} + \lambda\|\Theta\|_1\}$$

$\hat{\boldsymbol{S}}$ may not be positive semi-definite. Even though the problem is still convex, certain algorithms (like the blockwise coordinate descent algorithm) will fail. However, other algorithms that don't have positive semi-definite assumption on $\hat{\boldsymbol{S}}$ (like the two-metric projected Newtown method) can still work.

References I

[1] A. P. Dempster. "Covariance Selection". In: *Biometrics* 28.1 (1972), pp. 157–175. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2528966.

[2] Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. springer open, 2017.

[3] Han Liu et al. "High-dimensional semiparametric Gaussian copula graphical models". In: *The Annals of Statistics* 40.4 (2012), pp. 2293–2326.

[4] Rahul Mazumder and Trevor Hastie. "The graphical lasso: New insights and alternatives". In: *Electronic journal of statistics* 6 (2012), p. 2125.

References II

[5] T. P. Speed and H. T. Kiiveri. "Gaussian Markov Distributions over Finite Graphs". In: *The Annals of Statistics* 14.1 (1986), pp. 138–150. ISSN: 00905364. URL: http://www.jstor.org/stable/2241271.

[6] M.J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN: 9781108498029. URL: https://books.google.nl/books?id=IluHDwAAQBAJ.

[7] Ming Yuan and Yi Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (2007), pp. 19–35.