Notes on Covariance Estimation

Contents

1 Introduction 2

1.1 Example: Frequentist Estimation and Bayesian Estimation . . 2

2 Normal Distribution 3

2.1 Maximum Likelihood Estimation 4

3 Wishart distribution 6

3.1 Inverse-Wishart distribution 7

4 Schur Complement 7

5 Block Matrices 8

6 Concentration Graph Models 8

6.1 Some concepts from graph theory 8

6.2 Concentration Graph Models 9

6.2.1 (Negative) Log Likelihood 11

6.2.2 Iterative Proportional Fitting (IPF) 11

6.2.3 Decomposable graphs 17

6.2.4 Iterative Partial Minimization(IPM) 18

6.2.5 Application of IPM to maximizing $l^*(\Omega)$ 20

6.2.6 Bayesian Inference for Concentration Graph Models . . 24

6.2.7 The G-Wishart distribution 25

6.2.8 Sampling from the G-Wishart if G is decomposable . . 26

6.2.9 Block Gibbs-Sampling for G-Wishart if G is not decomposable 30

6.2.10 Other Sampling Mechanisms for the G-Wishart 32

7 Covariance Graph Models 32

7.1 Graphical Lasso 33

7.2 SPACE algorithm 34

7.3 CONCORD algorithm 38

1 Introduction

Consider the following problem: We want to estimate $\Sigma \in \mathbb{R}^{p\times p}$ where $\mathbf{Y}^1, ..., \mathbf{Y}^n \overset{iid}{\sim} (\mathbf{0}, \Sigma)$. An example of this could be 50 patients in a hospital who has 1000 genes recorded, where $n = 50$ and $p = 1000$. How would we estimate Σ using Frequentist or Bayesian paradigms?

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

where $\sigma_{ij} = \sigma_{ji}$. Thus we have to estimate $p + (p-1) + \cdots + 1 = \frac{p(p+1)}{2}$ parameters. Even if $n \approx p$, this is not a simple task. Frequentist approaches include putting an ℓ_1 penalty such as in the *lasso* regression, while Bayesian approaches involving finding an appropriate class of priors. We also consider estimating Ω but this problem is considered separately as inverting large matrices is usually not a good idea. The two models we consider are Covariance graph models (Σ is sparse) and Concentration graph models (Ω is sparse).

1.1 Example: Frequentist Estimation and Bayesian Estimation

Lemma 1. *Let $X^i \sim \mathcal{N}_p(0, \Sigma)$. Then $X^i(X^i)^T \sim \mathcal{W}_p(1, \Sigma)$.*

See Definiton 4.

1. Frequentist: $\hat{\Sigma}_{mle} = S = \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^T$. In this case, $nS \sim \mathcal{W}_p(n, \Sigma)$.

2. Bayesian: Note that as S is sufficient for Σ, and hence Ω, we can only consider $l(\Omega|S)$ instead of $l(\Omega|Data)$.

$$\begin{aligned} l(Data|\Sigma) = f(S) &\propto \frac{e^{-\frac{1}{2}tr(\Sigma^{-1}S)}}{|\Sigma|^{\frac{n}{2}}} \\ &= |\Omega|^{\frac{n}{2}} e^{-\frac{1}{2}tr(\Omega S)} \end{aligned}$$

Thus

$$l(Data|\Omega) = |\Omega|^{\frac{n}{2}} e^{-\frac{1}{2}tr(\Omega S)}.$$

Let $\mathbb{P}^+$ be the set of positive definite matrices and $\Lambda_0 \in \mathbb{P}^+$. If $\Omega \sim \mathcal{W}_p(\alpha + p + 1, \Lambda_0^{-1})$, then the pdf of Ω is

$$f(\Omega) \propto |\Omega|^{\frac{\alpha+p+1}{2}} e^{-\frac{1}{2}tr(\Lambda_0\Omega)}$$

Then

$$\begin{aligned} \pi(\Omega|Data) &\propto l(Data|\Omega)f(\Omega) \\ &= |\Omega|^{\frac{n+\alpha+p+1}{2}} e^{-\frac{1}{2}tr(n\Omega S+\Lambda_0\Omega)} \\ \implies \Omega|Data &\sim \mathcal{W}_p(n+\alpha+p+1,(nS+\Lambda_0)^{-1}) \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}[\Omega] &= (\alpha+p+1)(\Lambda_0)^{-1} \\ \mathbb{E}[\Omega^{-1}] &= \mathbb{E}[\Sigma] = \frac{\Lambda_0}{\alpha} \\ \mathbb{E}[\Omega|Data] &= (n+\alpha+p+1)(nS+\Lambda_0)^{-1} \\ \mathbb{E}[\Omega^{-1}|Data] &= \mathbb{E}[\Sigma] = \frac{nS+\Lambda_0}{n+\alpha} \\ &= \underbrace{\frac{n}{n+\alpha}S}_{\text{Linear function of Frequentist estimate}} + \underbrace{\frac{\alpha}{n+\alpha}(\frac{\Lambda_0}{\alpha})}_{\text{Linear function of Prior Mean}} \end{aligned}$$

Note that $\mathbb{E}[\Omega^{-1}|Data]$ is a convex combination of the frequentist estimate and the prior mean. This is a property of Diaconis-Ylvisaka priors (1979, Annals of Statistics).

2 Normal Distribution

Definition 1. *$X \sim \mathcal{N}(\mu, \sigma)$ if the density of X,*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2}(\frac{x-u}{\sigma})^2}.$$

Definition 2. *Let $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ such that Σ is positive definite. Then $\mathbf{X} \sim \mathcal{N}_p(\mu, \mathbf{\Sigma})$ if and only if*

$$\begin{aligned} f(\mathbf{x}) =& (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} e^{-(\mathbf{x}-\mu)^T\Sigma(\mathbf{x}-\mu)} \\ &\Leftrightarrow \forall \mathbf{a} \in \mathbb{R}^p, \mathbf{a}^T X \sim \mathcal{N}(\mathbf{a}^T\mu, \mathbf{a}^T\Sigma\mathbf{a}). \end{aligned}$$

Let $\mathbf{t} \in \mathbb{R}^p$*. Then the **characteristic function** of* $\mathbf{X}$ *is:*

$$\phi(\mathbf{t}) = e^{i\mathbf{t}^T\mu - \frac{1}{2}i\mathbf{t}^T\Sigma\mathbf{t}}.$$

Lemma 2. *Now suppose* $A \in \mathbb{R}^{m\times p}$*. Also suppose* $rank(A) = m$ *and* $\mathbf{b} \in \mathbb{R}^m$*. Then* $A\mathbf{X}+\mathbf{b} \in \mathbb{R}^m$ *and* $A\mathbf{X}+\mathbf{b} \sim \mathcal{N}(A\mu+\mathbf{b}, A\Sigma A^T)$*. Let* $\mathbf{X} = \left(\begin{smallmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{smallmatrix}\right) \sim \mathcal{N}(\left(\begin{smallmatrix}\mu_1\\ \mu_2\end{smallmatrix}\right), \left(\begin{smallmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{smallmatrix}\right))$*. Then*

1. $\mathbf{X}_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$
2. $\mathbf{X}_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$
3. $\mathbf{X}_2|\mathbf{X}_1 \sim \mathcal{N}_p(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$

Additionally,for normally distributed random vectors, $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X_2}$ *if and if* $\Sigma_{12} = \mathbf{0}$ *and* $\Sigma_{21} = \mathbf{0}$*. In such a case,* $\mathbf{X}_1 + \mathbf{X}_2 \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_{11} + \Sigma_{22})$*.*

2.1 Maximum Likelihood Estimation

Theorem 3. *Suppose* $\mathbf{X}^1, ..., \mathbf{X}^n \sim \mathcal{N}_p(\mu, \mathbf{\Sigma})$*. Then the maximum likelihood estimator, or mle, of* (μ, Σ) *is* $(\bar{\mathbf{X}}, S)$ *where* $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}^i$ *and* $S = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}^i - \bar{\mathbf{X}})(\mathbf{X}^i - \bar{\mathbf{X}})^T$*.*

Proof. First note that as $\sum_{i=1}^n (\mathbf{X}^i - \bar{\mathbf{X}})^T\Sigma^{-1}(\mathbf{X}^i - \bar{\mathbf{X}}) \in \mathbb{R}$,

$$\begin{aligned}
&\sum_{i=1}^n (\mathbf{X}^i - \bar{\mathbf{X}})^T\Sigma^{-1}(\mathbf{X}^i - \bar{\mathbf{X}})\\
=&tr(\sum_{i=1}^n (\mathbf{X}^i - \bar{\mathbf{X}})^T\Sigma^{-1}(\mathbf{X}^i - \bar{\mathbf{X}}))\\
=&tr(\sum_{i=1}^n (\mathbf{X}^i - \bar{\mathbf{X}})^T\Sigma^{-1}(\mathbf{X}^i - \bar{\mathbf{X}}))\\
=&tr(\sum_{i=1}^n (\mathbf{X}^i - \bar{\mathbf{X}})(\mathbf{X}^i - \bar{\mathbf{X}})^T\Sigma^{-1})\\
=&tr(nS\Sigma^{-1})\\
=&ntr(S\Sigma^{-1})
\end{aligned}$$

Thus,

$$\begin{aligned}
l(\mu,\Sigma) =& \log \prod_{i=1}^{n} f(\mathbf{x}^i) \\
=& -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}^i-\mu)^T\Sigma^{-1}(\mathbf{X}^i-\mu) \\
=& -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}^i-\mu)^T\Sigma^{-1}(\mathbf{X}^i-\mu) \\
=& -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}^i-\bar{\mathbf{X}}+\bar{\mathbf{X}}-\mu)^T\Sigma^{-1}(\mathbf{X}^i-\bar{\mathbf{X}}+\bar{\mathbf{X}}-\mu) \\
=& -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}^i-\bar{\mathbf{X}})^T\Sigma^{-1}(\mathbf{X}^i-\bar{\mathbf{X}}) - \frac{n}{2}(\bar{\mathbf{X}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{X}}-\mu) \\
=& -\frac{n}{2}\log|\Sigma| - \frac{n}{2}tr(\Sigma^{-1}S) - \frac{n}{2}(\bar{\mathbf{X}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{X}}-\mu)
\end{aligned}$$

It is clear at this point that for any value of Σ, the value of μ for which $l(\mu,\Sigma)$ is maximized is $\mu = \bar{\mathbf{X}}$ when $\frac{n}{2}(\bar{\mathbf{X}}-\mu)^T\Sigma^{-1}(\bar{\mathbf{X}}-\mu) = 0$. Thus, $\hat{\mu}_{mle} = \bar{\mathbf{X}}$. Now let

$$\begin{aligned}
F(\Sigma) =& -\log|\Sigma| - tr(\Sigma^{-1}S) \\
=& \log|\Sigma^{-1}| - tr(\Sigma^{-1}S^{\frac{1}{2}}S^{\frac{1}{2}}) + \log|S| - \log|S| \\
=& \log|\Sigma^{-1}S| - tr(\Sigma^{-1}S^{\frac{1}{2}}S^{\frac{1}{2}}) - \log|S| \\
=& \log|S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}| - tr(S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}) - \log|S|
\end{aligned}$$

In the last line we use the fact that $det(AB) = det(A)det(B) = det(B)det(A) = det(BA)$ and $tr(AB) = tr(BA)$. Now let $\lambda_1, ...\lambda_p$ be the eigenvalues of of $S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}$. Then recall that the trace of a matrix is the sum of the eigenvalues and the determinant is the product of the eigenvalues, which implies that

$$\begin{aligned}
&\log|S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}| - tr(S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}}) - \log|S| \\
=& \sum_{i=1}^{p}\log\lambda_i - \sum_{i=1}^{p}\lambda_i - \log|S|
\end{aligned}$$

As S is already known we can treat it like a constant. Then for each i, $\log\lambda_i - \lambda_i$ is minimized at $\lambda_i = 1$ because $\frac{d}{dx}(\log x - x) = \frac{1}{x} - 1 \overset{set}{=} 0 \implies x = 1$.

The second derivative, $-\frac{1}{x^2}$ is negative at $x = 1$ indicating a maximum point. Setting all the eigenvalues equal to 1 implies that $S^{\frac{1}{2}}\Sigma^{-1}S^{\frac{1}{2}} = I_{p\times p} \Leftrightarrow \Sigma = S$. Note that this result only holds provided $n > p$, which ensures that S is positive definite. □

Lemma 3. $(X-\mu)^T\Sigma^{-1}(X-\mu) \in \mathbb{R} \implies (X-\mu)^T\Sigma^{-1}(X-\mu) = tr((X-\mu)^T\Sigma^{-1}(X-\mu)) = tr(\Sigma^{-1}(X-\mu)(X-\mu)^T)$.

Lemma 4. *For $n > p$, S is almost surely a positive definite matrix (Eaton 2007, Das Gupta).*

Lemma 5. $\sum_{i=1}^{n}(X^i-\mu)^T\Sigma^{-1}(X^i-\mu) = ntr(\Sigma^{-1}S) + (\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)$.

Lemma 6. *Let $F : \mathbb{P}^+ \to \mathbb{R}$ such that $F(\Sigma) = -\log|\Sigma| - tr(\Sigma^{-1}S)$. If S is positive definite then $F(.)$ has a unique minimum, and this Rajaratnam occurs at $\Sigma = S$. The proof of this is almost identical to the proof that $\hat{\Sigma}_{mle} = S$.*

All these properties have been derived before, or their proofs are very similar to the proofs in the section on maximum likelihood.

3 Wishart distribution

Definition 4. *Suppose X is an $n \times p$ matrix, each row of which is independently drawn from a p-variate normal distribution with zero mean: i.e. $X = (X_{(1)}, ..., X_{(n)})^T$ where $X_{(i)}{=}(x_i^1, \ldots, x_i^p)^T \sim \mathcal{N}_p(0, V)$. Then the* ***Wishart distribution*** *is the probability distribution of the $p \times p$ random matrix, S, where*

$$\begin{aligned} S &= X^T X \\ &= (X_{(1)}, ..., X_{(n)})(X_{(1)}^T, ..., X_{(n)}^T) \\ &= \sum_{i=1}^{n} X_{(i)} X_{(i)}^T \end{aligned}$$

known as the scatter matrix.

One indicates that S has that probability distribution by writing $S \sim \mathcal{W}_p(V, n)$, or alternatively as $\mathcal{W}_p(n, V)$. The important thing to remember is that one of the parameters is an integer and the other is a positive definite matrix. The positive integer n is the number of degrees of freedom. Sometimes this is written $\mathcal{W}(V, p, n)$. For $n \geq p$ the matrix S is invertible with probability 1 if V is invertible. If $p = V = 1$ then this distribution is a chi-squared distribution with n degrees of freedom.

Definition 5. *The* ***density*** *of the Wishart Distribution is:*

$$f(s) = \frac{1}{2^{\frac{np}{2}}|V|^{\frac{n}{2}}\Gamma_p(\frac{n}{2})}|s|^{\frac{n-p-1}{2}}e^{\frac{-tr(V^{-1}s)}{2}}$$

where

$$\Gamma_p(a) = \pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma(a - \frac{i-1}{2})$$

In such a case, $\mathbb{E}(S) = nV$.

3.1 Inverse-Wishart distribution

The Wishart distribution is related to the Inverse-Wishart distribution, denoted by $\mathcal{W}_p^{-1}$, as follows:

Definition 6. *If $X \sim \mathcal{W}_p(V, n)$ and if we do the change of variables $C = X^{-1}$, then $C \sim \mathcal{W}_p^{-1}(V^{-1}, n)$.*

This relationship may be derived by noting that the absolute value of the Jacobian determinant of this change of variables is $|C|^{p+1}$. In such a case $\mathbb{E}(C) = \frac{V^{-1}}{n-p-1}$.

4 Schur Complement

Definition 7. *Let*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where A and D are square, and A is invertible. Then the ***Schur Complement*** *for A, denoted $M/A = D - CA^{-1}B$.*

In such a case,

$$det(M) = det(A)det(D - CA^{-1}B). \tag{1}$$

If instead D is invertible, the Schur complement for D, denoted $M/D = A - BD^{-1}C$.

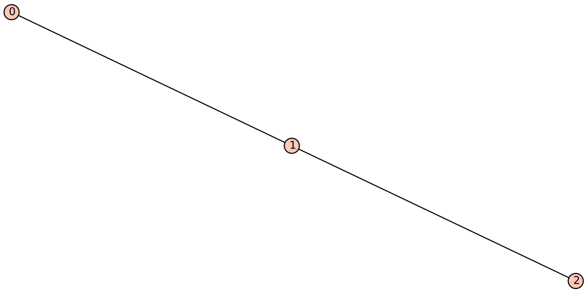


Figure 1: A graph with $V = \{0, 1, 2\}$ and $E = \{(0, 1), (1, 2)\}$.

5 Block Matrices

Theorem 8. *Let*

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \text{ and } Y = \begin{pmatrix} E & F \\ G & H \end{pmatrix}.$$

Then

$$tr(XY) = tr(AE + BG) + tr(CF + DH).$$

6 Concentration Graph Models

6.1 Some concepts from graph theory

Definition 9. *A **graph** G, is a collection of two 2 objects: V and E. We write $G = (V, E)$, where V is the set of vertices and $E \subset V \times V$.*

Figure 1 shows a simple graph with $V = \{0, 1, 2\}$ and $E = \{(0, 1), (1, 2)\}$.

Definition 10. *We say that u and v are **neighbors** if $(u, v) \in E$.*

In Figure 1, 0 and 1 are neighbors but 0 and 2 are not neighbors.

Definition 11. *A **p-cycle** us a collection of p distinct vertices, $u_1, ..., u_p$ such that the following properties hold:*

1. $(u_i, u_{i+1}) \in E, i = 1, ..., p.$

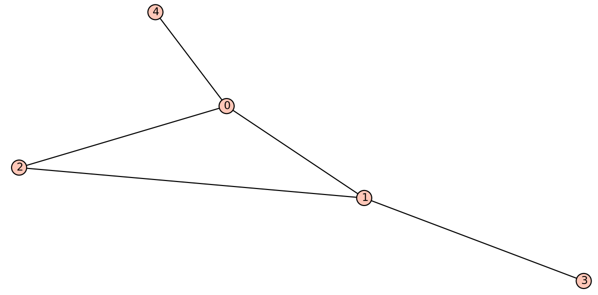


Figure 2: A graph with $V = \{0, 1, 2, 3, 4\}$ and $E = \{(0, 1), (0, 2), (0, 4), (1, 2), (1, 3)\}$.

2. $(u_p, u_1) \in E$.

Definition 12. *In the mathematical area of graph theory, a* ***clique*** *in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge, i.e.* V_0 *is a clique* $V_0 \subset V$ *such that* $\forall u \in V_0$ *and* $\forall v \in V_0$, $(u, v) \in E$. V_0 *is a* ***maximal clique*** *if*

1. V_0 *is a clique*

2. $\nexists \overline{V}$ *such that* $V_0 \subset \overline{V} \subset V$ *and* $\overline{V}$ *is a clique.*

In Figure 2, {0,1,2} and {1,3} are a maximal cliques, but {1,2} isn't.

6.2 Concentration Graph Models

Let $X^1, ...X^n \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$. We are interested in estimating $\Omega = \Sigma^{-1}$, where the Ω_{ij} are restricted to be 0.

Lemma 7. *(No distributional Assumptions) Let* $X \in \mathbb{R}^p$ *be a random vector and* $Cov(X) = \Sigma = \Omega^{-1}$. *Then* $\Omega_{ij} = Cov(X_i, X_j | X_k, k \neq i, k \neq j)$.

Lemma 8. *If we make the distributional assumption that* $X \sim \mathcal{N}_p(0, \Sigma)$, $\Omega_{ij} = 0 \Leftrightarrow X_i | X_k \perp\!\!\!\perp X_j | X_k, k \neq i, k \neq j$.

Example It makes sense to look at conditional covariances as in many cases it turns out that variables that are marginally dependent are conditionally independent. Consider income, race and crime. Marginally, it seems that crime and race are dependent but after conditioning on income, crime and race are independent.

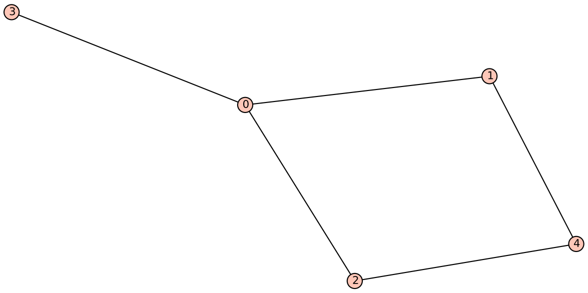


Figure 3: A graph with $V = \{0,1,2,3,4\}$ and $E = \{(0,1),(0,2),(0,3),(1,4),(2,4)\}$.

Model Corresponding to Graph, $G = (V, E)$**:** Let $\Omega \in \mathbb{P}_G := \{\ A|A \in \mathbb{P}^+$ and $A_{ij} = 0$ whenever $(i,j) \notin E\}$ and suppose that the number of elements in $V, |V| = p$. As an example consider the graph, shown in Figure 3,

$$\begin{aligned} G_1 &= (V, E), \\ \text{where } V &= \{0,1,2,3,4\} \\ \text{and } E &= \{(0,1),(0,2),(0,3),(1,4),(2,4)\}. \end{aligned}$$

Then $\Omega \in \mathbb{P}_{G_1}$ and the corresponding concentration matrix for this graph is

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & 0 \\ \omega_{21} & \omega_{22} & 0 & 0 & \omega_{25} \\ \omega_{31} & 0 & \omega_{33} & 0 & \omega_{35} \\ \omega_{41} & 0 & 0 & \omega_{45} & 0 \\ 0 & \omega_{52} & \omega_{53} & 0 & \omega_{55} \end{pmatrix}.$$

Suppose $n = 2p > p$. Then $\hat{\Omega} = S^{-1}$ is a valid estimate as S^{-1} exists. However, we cannot put 0s into S^{-1} arbitrarily as we need to preserve the positive definite structure to have a valid estimate.

Lemma 9. *If A is positive definite and we construct*

$$A_G = \begin{cases} A_{ij} & \textit{if } (i,j) \in E \\ 0 & \textit{if } (i,j) \notin E \end{cases}$$

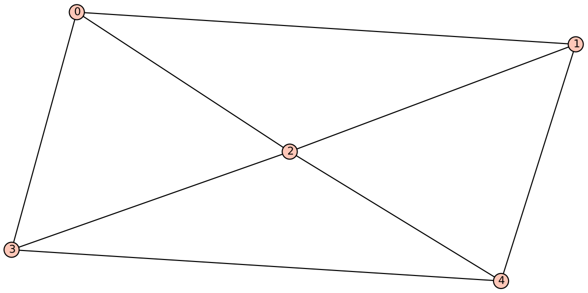


Figure 4: A graph with $V = \{0, 1, 2, 3, 4\}$ and $E = \{(0, 1), (0, 2), (0, 3), (1, 2), (1, 4), (2, 3), (3, 4)\}$.

Then A is not positive definite in general. (Under some assumptions A_G may be positive definite asymptotically but that still doesn't give us an estimator for a fixed n and p.)

6.2.1 (Negative) Log Likelihood

If $X^1, ..., X^n \sim \mathcal{N}_p(0, \Sigma = \Omega^{-1}), G = (V, E), |V| = p, \Omega \in \mathbb{P}_G$, then the negative log likelihood is,

$$l(\Omega) = c + \frac{n}{2} tr(\Omega S) - \frac{n}{2} \log|\Omega|$$

where $\Omega \in \mathbb{P}_G$ and c is a constant. $l(\Omega)$ has a unique global minimum if $n > max\{|C_1|, ..., |C_n|\}$ where $C_1, ..., C_n$ denotes the cliques of G. In Figure 4, $p = 5$ and the cliques are $C_1 = 0, 1, 2$,$C_2 = 0, 2, 3$,$C_3 = 2, 3, 4$,$C_4 = 1, 2, 4$. Thus $max|C_i| = 3$. Hence for a unique maximum likelihood estimator to exist we need $n > 3$. Now for $\Omega \in \mathbb{P}_G$, consider $l^*(\Omega) = tr(\Omega S) - \log|\Omega|$, which has a minimum at the same Ω as $l(\Omega)$. In general there is no closed form for the global minimum. Hence we need to use iterative minimization techniques.

6.2.2 Iterative Proportional Fitting (IPF)

Speed and Kiveri (1986) came up with the following algorithm:

1. Start with an initial estimate $\Omega^0 \in \mathbb{P}_G$.

2. Set $\Omega^{(r,0)} = \Omega^0$.

3. Repeat the following for $i = 1, ..., k$ where k is the number of cliques and the vertex set, $V = C_i \cup \bar{C}_i$ and $\bar{C}_i = V \setminus C_i$:

   - Set $\Omega^{(r,i)} = \Omega^{i-1}$
   - $\Omega^{(r,i)} = \arg\min\{tr(AS) - \log|A|\}$ where

   $$A : \begin{cases} (A^{-1})_{C_i\bar{C}_i} = \Sigma^{(r,i-1)}_{C_i\bar{C}_i} \\ (A^{-1})_{\bar{C}_i\bar{C}_i} = \Sigma^{(r,i-1)}_{\bar{C}_i\bar{C}_i} \end{cases}$$

   More details on this step below.

4. If $\|\Omega^{(r,k)} - \Omega^{(r,0)}\| < tol$, stop. Else set $\Omega^{(r+1,0)} = \Omega^{(r,k)}$ and go back to Step 3.

To minimize the function in Step 3, first we permute the rows and columns of Σ to get

$$\Sigma = \begin{pmatrix} \Sigma_{C_1C_1} & \Sigma_{C_1\bar{C}_1} \\ \Sigma_{C_1\bar{C}_1} & \Sigma_{\bar{C}_1\bar{C}_1} \end{pmatrix}$$

Then let

$$\begin{aligned} \Omega &= \Sigma^{-1} \\ &= \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \end{aligned}$$

where

$$\begin{aligned} \Omega_{11} &= (\Sigma_{C_1C_1} - \Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1})^{-1} \\ \Omega_{12} &= -\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1} \\ \Omega_{21} &= -\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{\bar{C}_1C_1}\Omega_{11} \\ \Omega_{22} &= \Sigma^{-1}_{\bar{C}_1\bar{C}_1} + \Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1}\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1} \end{aligned}$$

and

$$S = \begin{pmatrix} S_{C_1C_1} & S_{C_1\bar{C}_1} \\ S_{\bar{C}_1C_1} & S_{\bar{C}_1\bar{C}_1} \end{pmatrix}$$

Therefore,

$$\begin{aligned}
tr(\Omega S) &= tr(\Sigma^{-1}S) \\
&= tr\Big(\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \times \begin{pmatrix} S_{C_1C_1} & S_{C_1\bar{C}_1} \\ S_{\bar{C}_1C_1} & S_{\bar{C}_1\bar{C}_1} \end{pmatrix} \Big) \\
&= tr\Big(\begin{pmatrix} \Omega_{11}S_{C_1C_1} + \Omega_{12}S_{\bar{C}_1C_1} & \Omega_{11}S_{C_1\bar{C}_1} + \Omega_{12}S_{\bar{C}_1\bar{C}_1} \\ \Omega_{21}S_{C_1C_1} + \Omega_{22}S_{\bar{C}_1C_1} & \Omega_{21}S_{C_1\bar{C}_1} + \Omega_{22}S_{\bar{C}_1\bar{C}_1} \end{pmatrix} \Big) \\
&= tr(\Omega_{11}S_{C_1C_1} + \Omega_{12}S_{\bar{C}_1C_1}) + tr(\Omega_{21}S_{C_1\bar{C}_1} + \Omega_{22}S_{\bar{C}_1\bar{C}_1}) \\
&= tr(\Omega_{11}S_{C_1C_1} + 2\Omega_{12}S_{\bar{C}_1C_1} + \Omega_{22}S_{\bar{C}_1\bar{C}_1}) \\
&= tr(\Omega_{11}S_{C_1C_1} + -2\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1C_1} + \Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1} \\
&\qquad + \Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1}\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1})
\end{aligned}$$

Thus

$$\begin{aligned}
tr(\Omega S) &= tr(\Sigma^{-1}S) \\
&= tr(\Omega_{11}S_{C_1C_1}) + tr(-2\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1C_1}) + tr(\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1}) \\
&\qquad + tr(\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1}\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1}) \\
&= tr(\Omega_{11}S_{C_1C_1}) + tr(-2\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1C_1}) + tr(\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1}) \\
&\qquad + tr(\Omega_{11}\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1}) \\
&= tr(\Omega_{11}(S_{C_1C_1} - 2\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1C_1} + \Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1})) \tag{2} \\
&\qquad + tr(\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1})
\end{aligned}$$

For the second term: $\log|\Omega|$, recall from Equation 1 that if

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

and D is invertible, then $|M| = |D||A - BD^{-1}C|$. Thus,

$$\begin{aligned}
|\Sigma| &= |\Sigma_{\bar{C}_1\bar{C}_1}||\underbrace{\Sigma_{C_1C_1} - \Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1}|}_{\Omega_{11}^{-1}} \\
&= |\Sigma_{\bar{C}_1\bar{C}_1}||\Omega_{11}^{-1}| \\
&= \frac{|\Sigma_{\bar{C}_1\bar{C}_1}|}{|\Omega_{11}|} \\
\implies \log|\Omega| &= \log|\Sigma^{-1}| \\
&= \log(\frac{1}{|\Sigma|}) \\
&= \log\left(\frac{|\Omega_{11}|}{|\Sigma_{\bar{C}_1\bar{C}_1}|}\right) \\
&= \log|\Omega_{11}| - \log|\Sigma_{\bar{C}_1\bar{C}_1}|
\end{aligned}$$

As C_1 is a clique, all the nodes in C_1 are connected to every other node in C_1 and hence there are no 0's in Ω_{11}. Thus Ω_{11} is positive definite with no constraints. The idea in IPF is to maximize $l^*(\Omega)$ over C_1 while hold everything else constant, which in this case is $\Sigma_{C_1\bar{C}_1}$ and $\Sigma_{\bar{C}_1\bar{C}_1}$.

$$\begin{aligned}
l^*(\Omega) &= tr(\Omega S) - \log|\Omega| \\
&= tr(\Sigma^{-1}S) - \log|\Omega| \\
&= tr(\Omega_{11}(S_{C_1C_1} - 2\Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1C_1} + \Sigma_{C_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}S_{\bar{C}_1\bar{C}_1}\Sigma^{-1}_{\bar{C}_1\bar{C}_1}\Sigma_{C_1\bar{C}_1})) \\
&\quad + \log|\Omega_{11}| + \text{terms not depending on } \Omega_{11}
\end{aligned}$$

To maximize with respect to Ω_{11} we can simply take the derivative and set it equal to 0. As C_1 is a clique, $S_{C_1C_1}$ is clearly positive definite as $n > |C_1|$. It can also be shown that S is positive definite in such a case. (Need a reference/proof of this fact).

Lemma 10. *If $n > max\{C_1, ...C_k\}$,i.e. the sample size is bigger than the largest clique size, then $l^*(\Omega)$ is strictly convex. Lauritzen(1996) has conditions for convergence of the partial minimization algorithm under convexity of $l^*(\Omega)$.*

For non-convex $l^*(\Omega)$ look at Drton and Elder (2006).

Definition 13. *let $G = (V, E)$ be a graph. Then $G_1 = (V_1, E_1)$ is a **subgraph** of G if $V_1 \subset V$ and $E_1 \subset E$. In addition, if $(u, v) \in E_1$ whenever $u \in V_1 \& v \in V_1$, then G_1 is an **induced subgraph** of G.*

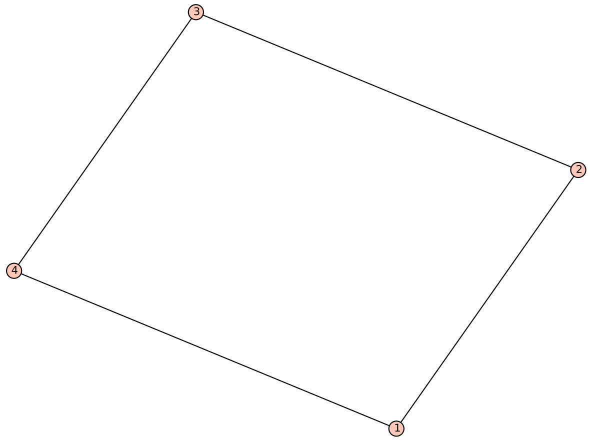
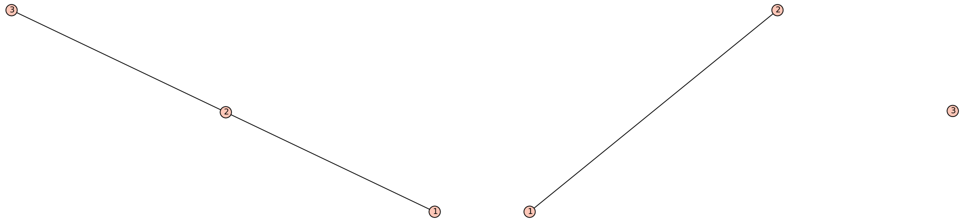


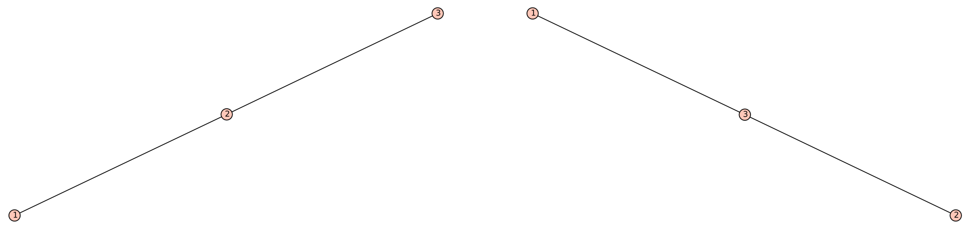
Figure 5: A graph with $V = \{1, 2, 3, 4\}$, $E = \{(4, 1), (1, 2), (2, 3), (3, 4)\}$.



(a) An induced graph with $V_1 = \{1, 2, 3\} \subset V$, $E_1 = \{(2, 3), (1, 2)\} \subset E$.

(b) A subgraph with $V_1 = \{1, 2, 3\} \subset V$, $E_1 = \{(1, 2)\} \subset E$.

Figure 6: Subgraphs



(a) Not a perfect elimination ordering of the vertices

(b) A perfect elimination ordering of the vertices

Figure 7: Perfect Elimination Orderings

Definition 14. *(We need this concept in Definition 2 of decomposable graphs) Let $G = (V, E)$ be a graph. An* ***ordering of the vertices*** *is a bijection,σ from V to the set $\{1, 2, ..., |V|\}$. Then the* ***ordered graph*** *$G_\sigma = (V_\sigma, E_\sigma)$ where $V_\sigma = \{1, 2, ..., |V|\}$ and $(i, j) \in E_\sigma \iff (\sigma^{-1}(i), \sigma^{-1}(j)) \in E$. An ordering σ is defined to be a* ***perfect elimination ordering*** *if $\nexists i > j > k$ such that $(i, j) \notin E_\sigma$ but $(j, k) \in E_\sigma$ and $(i, k) \in E_\sigma$.*

Consider the graph in Figure 6a. If

$$\sigma_1 = \begin{pmatrix} 0 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix}$$

then $3 > 2 > 1$ and $(2, 3) \notin E_\sigma$ but $(2, 1) \in E_\sigma$ and $(3, 1) \in E_\sigma$. Thus σ_1 is not a perfect elimination ordering. However,

$$\sigma_2 = \begin{pmatrix} 0 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}$$

is a perfect elimination ordering scheme.

Definition 15. *(We need this concept in Definition 3 of decomposable graphs.) If Ω is a positive definite matrix, then $\exists$ a unique pair (L, D) such that*

1. *$\Omega = LDL^T$ is the* ***modified Cholesky decomposition*** *of Ω*

2. *L is a lower triangular matrix with 1's on the diagonals*

3. *D is a postive diagonal matrix (i.e. $d_{ii} > 0 \forall i$)*

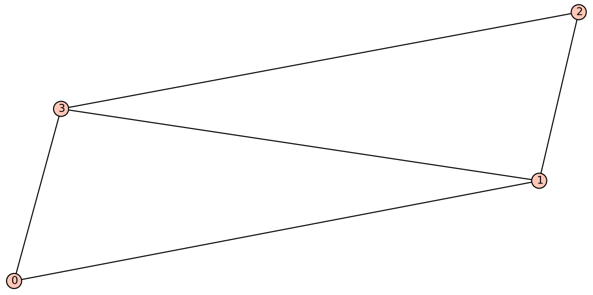


Figure 8: A graph with $V = \{0, 1, 2, 3\}$, $E = \{(0, 1), (0, 3), (1, 2), (1, 3), (2, 3)\}$.

6.2.3 Decomposable graphs

There are many equivalent definitions of decomposable graphs.

Definition 16. *A graph $G = (V, E)$ is **decomposable** if and only if it does not contain a cycle of length greater than equal to 4.*

Figure 5 is not decomposable, but Figure 8 is.

Definition 17. *A graph is **decomposable** if and only if it has a perfect elimination ordering. Thus if we can find a perfect elimination ordering then it means that the graph is decomposable.*

Definition 18. *A graph $G = (V, E)$ is **decomposable** if and only if there exists and ordering, σ, of the vertices such that if $\Sigma = LDL^T$ is the modified Cholesky decomposition corresponding to this ordering, then for $i > j, L_{ij} = 0 \iff \Sigma_{ij} = 0 \iff (i, j) \notin E_\sigma$. Note that the order is of utmost importance due to uniqueness. If the order is changed, then we get a new Cholesky decomposition.*

Definition 19. *A graph is **decomposable** if and only if there is no chorldless cycle of length greater than or equal to 4 as an induced subgraph.*

Definition 20. *Let $G = (V, E)$ be a decomposable graph. Then there exists an ordering of the maximal cliques $C_1, ..., C_k$ such that for every $2 \leq j \leq k$*

$$R_j = C_j \cap (\cup_{l=1}^{j-1} C_l)$$

1. *$R_j \subset C_i textforsome 1 \leq i \leq j-1$.*

2. *R_j is called the j-th **minimal separator***

The above definition simply states that we can order the cliques so that for any clique, say C_h, in that given ordering we can find some previous clique,$C_i, 1 \leq i \leq h$, that contains the intersection of the the clique, C_h with all previous cliques, $C_1, ..., C_{h-1}$.

6.2.4 Iterative Partial Minimization(IPM)

We first consider coordinate wise minimization, which is a special case of IPM. Suppose we are interested in minimizing a function, $f(x), x = (x_1, ..., x_p) \in \mathcal{X}$. Note that $x \mapsto (x_i, x_{-i})$ is a bijection. Coordinate-wise minimization consists of repeating the following steps for $i = 1, ..., p$

1. Minimize f with respect to x_i holding the other blocks constant.

In IPM our main objective is to minimize $f(x), x \in \mathcal{X}$. Let $x \mapsto (y^i, y^{-i})$ be a bijection. For example, let $x = (x_1, x_2)$, $y^i = x_1 + x_2$ and $y^{-i} = x_1 - x_2$. Then $x \mapsto (y^i, y^{-i})$ is a bijection as given any x we can find y^i and y^{-i} and vice versa. Now, the main idea of IPM is to minimize f with respect to y^i holding y^{-i} constant, and then with respect to y^{-i} holding y^i constant. Recall that in IPF our goal is to minimize $l^*(\Omega)$. Earlier we had mentioned that IPF is a special case of IPM. Consider partitioning Ω in the following manner:

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

Now $\Omega_{12} = \Omega_{21}'$, thus $\Omega \mapsto (\Omega_{11}, (\Omega_{21}, \Omega_{22}))$ is a bijection. Thus according to IPM maximizing over Ω_{11} holding $(\Omega_{21}, \Omega_{22})$ constant and then maximizing over $(\Omega_{21}, \Omega_{22})$ while holding Ω_{11} constant would be a valid approach. However, ensuring positive definiteness of Ω is difficult if we approach the problem in this manner. As a result we consider the bijection: $\Omega \mapsto (\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}, (\Omega_{21}, \Omega_{22}))$.

Theorem 21. *(Iterated Partial Maximization) If*

1. *$L : \Theta \to \mathbb{R}$ is continuous and Θ is compact*

2. *$\forall\theta^* \in \Theta, \exists$ section,$\Theta_i(\theta^*), i = 1, ..., k$ in Θ in such a way that L is globally maximized at θ^* if and only if L is maximized over all of the sections.*

3. *The operations of maximizing L over the sections is continuous and well defined, i.e. there are continuous transformations T_i of Θ into itself such that if $\theta \in \Theta_i(\theta^*)$ for $i = 1, ..., k$*

   $$L\{T_i(\theta^*)\} > L(\theta), \quad \theta \neq T_i(\theta^*)$$

   In other words $T_i(\theta^*)$ is the uniquely determined point where L is maximized over the section $\Theta_i(\theta^*)$. Now let θ_0 be arbitrary and define recursively

   $$\theta_{n+1} = T_1...T_k(\theta_n), \quad n \geq 0.$$

4. *$L(\theta)$ is uniquely maximized at $\hat{\theta}$.*

Then $\theta_n \to \hat{\theta}$.

Proof. Since Θ is compact, the sequence (θ_n) has a convergent subsequence (θ_{n_l}) such that as $l \to \infty$, $(\theta_{n_l}) \to \theta^* \in \Theta$. We need to show that $\hat{\theta} = \theta^*$. Let $S = T_1...T_k$, that is, $\theta_{n+1} = S(\theta_n)$. Since each T-operation is a partial maximization, that is $L\{T_i(\theta^*)\} > L(\theta), \forall\theta \neq T_i(\theta^*)$, $L(\theta_n)$ must be non-decreasing in n. Thus $n + 1 > n \implies L(\theta_{n+1}) \geq L(\theta_n)$. Hence as $n_{l+1} \geq n_l + 1 > n_l$ we have $L(\theta_{n_{l+1}}) \geq L(\theta_{n_l})$. Also, limits are preserved by continuity. Thus,

$$\begin{aligned}
L\{S(\theta^*)\} &= \lim_{l\to\infty} L\{S(\theta_{n_l})\} \quad \text{by 1,2 and } l \to \infty, (\theta_{n_l}) \to \theta^* \in \Theta \\
&\leq \lim_{l\to\infty} L(\theta_{n_{l+1}}) \\
&= L(\theta^*) \\
&\leq L\{T_k(\theta^*)\} \quad \text{as each } T_i \text{ is a partial maximization} \\
&\ \ \vdots \\
&\leq L\{T_1...T_k(\theta^*)\} \\
&\leq L\{S(\theta^*)\}
\end{aligned}$$

Thus there must be equality at every step. As the partial maxima are unique, we also have that

$$\theta^* = T_k(\theta^*) = ... = T_1(\theta^*).$$

Finally since the global maximum was uniquely determined by maximizing L over all sections, the proof is complete. □

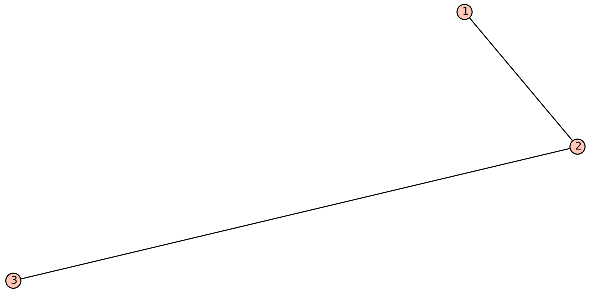


Figure 9: A graph with $V = \{1, 2, 3\}$, $E = \{(1, 2), (2, 3)\}$.

6.2.5 Application of IPM to maximizing $l^*(\Omega)$

Let

$$\begin{aligned}
X^1, \ldots X^n &\overset{iid}{\sim} \mathcal{N}_p(0, \Sigma = \Omega^{-1}) \\
\Omega \in \mathbb{P}_G &= \{A \in \mathbb{P}^+ | A_{ij} = 0 \iff (i, j) \notin E\} \\
\mathbb{P}^+ &= \{p \times p \text{ positive definite matices }\} \\
\mathcal{L}_G &= \{L | L \text{ is lower triangular;} \\
&\qquad L_{ii} = 1 \forall i = 1, \ldots p; \\
&\qquad i > j, (i, j) \notin E \implies L_{ij} = 0\} \\
\mathcal{D} &= \{D | D \text{ is diagonal with} D_{ii} > 0\}
\end{aligned}$$

Consider the graph in Figure 9. The corresponding $L \in \mathcal{L}_G$ is

$$L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ 0 & l_{32} & 0 \end{pmatrix}$$

where $l_{31} = 0$ as $(3, 1) \notin E$. Now recall that from Definition 18 that G is decomposable if and only if there is a perfect vertex elimination scheme if and only if the ordering of the vertices for that perfect vertex elimination scheme implies $\Omega = LDL^T$ and $(i, j) \notin E \implies L_{ij} = 0$. In addition, $\Omega \mapsto (L, D)$ such that $\Omega = LDL^T$ is a bijection from $\mathbb{P}_G$ to $\mathcal{L}_G \times \mathcal{D}$. This is because, once an ordering of the vertices has been fixed, the cholesky decomposition is unique. In general, a positive definite matrix has a unique cholesky decomposition, but there may be several perfect elimination orderings. Thus given and Ω we can find a unique L and a unique D and vice versa. As a result, using the concepts from IPM, to minimize $l^*(\Omega)$ we can minimize $l^*(L, D)$ instead. Note that

$$\begin{aligned}
l^*(\Omega) &= tr(\Omega S) - \log|\Omega| \\
\implies l^*(L,D) &= tr(LDL^T S) - \log|LDL^T| \\
&= tr(DL^T SL) \quad \{tr(AB) = tr(BA)\} \\
&\quad\ - \log|L||D||L^T| \quad \{det(ABC) = det(A)det(B)det(C)\} \\
&= tr(DL^T SL) - \log|D| \quad \{det(L) = \prod_{i=1}^{p} L_{ii} = 1\} \\
&= tr(DL^T SL) - \log\prod_{I=1}^{p} D_{ii} \quad \{det(D) = \prod_{i=1}^{p} D_{ii} = 1\} \\
&= \sum_{i=1}^{p}(D_{ii}(L_{.i}^T SL_{.i}) - logD_{ii})
\end{aligned}$$

In the final step we have used the following facts

1. Pre-multiplying a matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{pmatrix}$$

by a diagonal matrix

$$D = \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_{pp} \end{pmatrix}$$

leads to multiplying the i-th row of A by d_{ii}, i.e.

$$AD = \begin{pmatrix} d_{11}a_{11} & d_{11}a_{12} & \dots & d_{11}a_{1p} \\ d_{22}a_{21} & d_{22}a_{22} & \dots & d_{22}a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{pp}a_{p1} & d_{pp}a_{p2} & \dots & d_{pp}a_{pp} \end{pmatrix}$$

2. $(SL)_{ij} = S_{i.}L_{.j} \implies (SL)_{.j} = SL_{.j}$

$$\begin{aligned}(SL)_{1j} &= S_{1.}L_{.j}\\ (SL)_{2j} &= S_{2.}L_{.j}\\ &\cdots\\ (SL)_{pj} &= S_{p.}L_{.j}\end{aligned}$$

3. The ii-th entry of L^TSL:

$$(L^TSL)_{ii} = (L^T)_{i.}(SL)_{.i} = (L_{.i})^TSL_{.i}$$

Now note that for each i we can minimize each term in the sum independently of $j \neq i, j = 1, ..., p$. For example, if $i = 1$, we can minimize, $(D_{11}(L_{.1}^TSL_{.1}) - \log D_{11}$ and then move on to $i = 2$. In this regard note that

$$L_{.i}^TSL_{.i} = \begin{pmatrix}1 & x_i^T\end{pmatrix}\begin{pmatrix}S_{ii} & S_{.i}^{>}\\ S_{.i}^{>} & S^{>i}\end{pmatrix}\begin{pmatrix}1\\ x_i\end{pmatrix} \tag{3}$$

where

- $x_i = (L)_{ji}, j > i, (i,j) \in E$, i.e. the elements of the i-th columns of L that lie below the diagonal such that there is an edge between i and j.
- $S^{>i} = (S_{jk})_{j,k>i}, (i,j) \in E, (i,k) \in E$, i.e. the submatrix of S from the $(i+1)$-th row and $(i+1)$-th column through to the p-th row and p-th column such that there is an edge between i and j and i and k.
- $S_{.i}^{>} = (S_{ji})_{j>i}, (i,j) \in E$, i.e. the vector from the i-th column of S that lie below the diagonal, S_{ii} such that there is an edge between i and j.

For each $i = 1, ...p$, first we minimize Equation 3 with respect to x_i. The solution turns out to be

$$\hat{x}_i = -(S^{>i})^{-1}S_{.i}^{>}.$$

Then we minimize $(D_{ii}(S_{ii} - (S_{.i}^{>})^T(S^{>i})^{-1}(S_{.i}^{>})) - \log D_{ii}$ with respect to D_{ii}. This turns out to be

$$\hat{D}_{ii} = \frac{1}{S_{ii} - (S_{.i}^{>})^T(S^{>i})^{-1}(S_{.i}^{>})}.$$

Then we can estimate Ω using

$$\hat{\Omega} = \hat{L}\hat{D}\hat{L}^T \tag{4}$$

Now suppose that $C_1, ..., C_k$ is and ordering of the maximal cliques of $G = (V, E)$ and let $R_2, ..., R_k$ be the minimal separators as in Definiton 20. Then

$$\hat{\Omega} = \sum_{i=1}^{k} [(S_{C_i})^{-1}]^0 - \sum_{i=2}^{k} [(S_{R_i})^{-1}]^0 \tag{5}$$

where for $A \subset V$

$$([(S_A)^{-1}]^0)_{kl} = \begin{cases} S_A^{-1} & \text{if } k \in A, l \in A \\ 0 & \text{if } k \notin A \text{ or } l \notin A \end{cases}$$

Example: Let

$$\Omega = \begin{pmatrix} 0.80 & 0.37 & 0.00 & 0.31 \\ 0.37 & 1.37 & 0.58 & 0.39 \\ 0.00 & 0.58 & 0.32 & 0.16 \\ 0.31 & 0.39 & 0.16 & 0.69 \end{pmatrix}$$

For simplicity suppose we have an exact estimate of S:

$$S = \Omega^{-1} = \begin{pmatrix} 3.33 & -3.57 & 7.05 & -1.10 \\ -3.57 & 7.18 & -13.36 & 0.62 \\ 7.05 & -13.36 & 28.52 & -2.19 \\ -1.10 & 0.62 & -2.19 & 2.09 \end{pmatrix}$$

Then

$$S_{C_1} = \begin{pmatrix} 3.33 & -3.57 & -1.10 \\ -3.57 & 7.18 & 0.62 \\ -1.10 & 0.62 & 2.09 \end{pmatrix} \text{ and } (S_{C_1})^{-1} = \begin{pmatrix} 0.80 & 0.37 & 0.31 \\ 0.37 & 0.32 & 0.10 \\ 0.31 & 0.10 & 0.61 \end{pmatrix}$$

which implies that

$$[(S_{C_1})^{-1}]^0 = \begin{pmatrix} 0.80 & 0.37 & 0.00 & 0.31 \\ 0.37 & 0.32 & 0.00 & 0.10 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.31 & 0.10 & 0.00 & 0.61 \end{pmatrix}.$$

Similarly

$$\implies [(S_{C_2})^{-1}]^0 = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.19 & 0.58 & 0.25 \\ 0.00 & 0.58 & 0.32 & 0.16 \\ 0.00 & 0.25 & 0.16 & 0.57 \end{pmatrix}$$

and

$$[(S_{R_2})^{-1}]^0 = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.14 & 0.00 & -0.04 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & -0.04 & 0.00 & 0.49 \end{pmatrix}.$$

Finally,

$$\hat{\Omega} = [(S_{C_1})^{-1}]^0 + [(S_{C_2})^{-1}]^0 - [(S_{R_2})^{-1}]^0 = \Omega.$$

6.2.6 Bayesian Inference for Concentration Graph Models

Let

$$f_\theta(x) = e^{x'\theta - \kappa(\theta)} h(x)$$

where $\theta \in \tilde{\Theta} \subset \mathbb{R}^d$. Let $\tilde{\pi}_{n_0,x0}(\theta)$ denote a family of prior distributions for the natural parameter θ with $n_0 \in \mathbb{R}, x_0 \in \mathbb{R}^d$ given by

$$\tilde{\pi}_{n_0,x0}(\theta) = e^{n_0 x_0' \theta - n_0 \kappa(\theta)}.$$

If $\tilde{\pi}_{n_0,x_0}(\theta)$ can be normalized to define a valid probability distribution say $\pi_{n_0,x_0}(\theta)$, then it is a valid **conjugate prior** to the Natural Exponential Family (NEF).

Lemma 11. *Furthermore, if $X_1, ...X_n \overset{iid}{\sim} f_\theta(x), \theta \in \tilde{\Theta}$, then*

1. *The **posterior density** is given by $\pi_{n_0+n, \frac{n_0 x_0 + n\bar{X}}{n_0+n}}(\theta)$*

2. *The **posterior expectation** of $\frac{\partial \kappa(\theta)}{\partial \theta}$, $\mathbb{E}[\frac{\partial \kappa(\theta)}{\partial \theta}|X_1, ..., X_n] = \frac{n_0 x_0 + n\bar{X}}{n_0+n}$.*

3. *If, in addition, $\pi(\theta)$ is any prior such that it is not concentrated at a single point and, $\mathbb{E}[\frac{\partial \kappa(\theta)}{\partial \theta}|X] = aX + b$ for some constants a, b, then $a \neq 0$ and the density is necessarily of the form*

$$\pi(\theta) = c e^{\frac{1}{a} b\theta - \frac{1}{a}(1-a)\kappa(\theta)}$$

In other words, it is a Diaconis-Ylvisaker(DY) prior.

Example Suppose $X_1, ...X_n \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma), \Sigma = \Omega^{-1}$ and $\Omega \sim \mathcal{W}_p(m, \Lambda_0^{-1})$. Thus,

$$f(\Omega) \propto e^{-\frac{1}{2}tr(\Lambda_0\Omega)+\frac{m-p-1}{2}\log|\Omega|}$$

Now $nS = \sum_{i=1}^n X_i X_i^T \sim \mathcal{W}_p(n, \Sigma)$. Keeping only the terms containing Ω implies that

$$\begin{aligned} f(nS) &\propto e^{-\frac{1}{2}tr(\Sigma^{-1}nS)+\frac{n}{2}\log|\Sigma^{-1}|} \\ &= e^{-\frac{1}{2}tr(\Omega nS)+\frac{n}{2}\log|\Omega|} \\ \implies f(\Omega|S) &\propto e^{-\frac{1}{2}tr(\Lambda_0\Omega)+\frac{m-p-1}{2}\log|\Omega|-\frac{1}{2}tr(\Omega nS)+\frac{n}{2}\log|\Omega|} \\ &= e^{-\frac{1}{2}tr(\Omega(\Lambda_0+nS))+\frac{m+n-p-1}{2}\log|\Omega|} \end{aligned}$$

and by Lemma 11 take $\pi_{n_0,x_0}(\theta) = \mathcal{W}_p(m, \Lambda_0^{-1})$ where $n_0 = m - p - 1$ and $x_0 = \frac{\Lambda_0}{m-p-1}$, then $n_0 + n = m + n - p - 1$ and $\frac{n_0 x_0 + n\bar{X}}{n_0+n} = \frac{\Lambda_0+nS}{m+n-p-1}$ or directly,

$$\begin{aligned} \Omega|S &\sim \mathcal{W}_p(n+m, (nS+\Lambda_0)^{-1}) \\ \mathbb{E}[\Sigma] = \mathbb{E}[\Omega^{-1}] &= \frac{\Lambda_0}{m-p-1} \\ \text{and } \mathbb{E}[\Sigma|S] = \mathbb{E}[\Omega^{-1}|S] &= \frac{nS+\Lambda_0}{n+m-p-1} \end{aligned}$$

6.2.7 The G-Wishart distribution

Now suppose $X^1, ...X^n \overset{iid}{\sim} \mathcal{N}_p(0, \Omega^{-1})$. Then

$$l(\Omega) = -\frac{n}{2}tr(\Omega S) + \frac{n}{2}\log|\Omega|$$

where $\Omega \in \mathbb{P}_G$. A problem that could arise with $\Omega \in \mathbb{P}_G$ is that all the nice properties of natural exponential families when $\Omega \in \mathbb{P}^+$ may not be retained. If $\Omega \in \mathbb{P}^+$, then it is a natural exponential family , but restricting some of the entries to 0 could possibly violate the properties of NEFs. Fortunately, restricting some of the entries to be 0 does not affect the properties. If $\Omega \in \mathbb{P}_G$, then

$$\begin{aligned} tr(\Omega S) &= \sum_{i=1}^p \sum_{j=1}^p \Omega_{ij} S_{ij} = \sum_{(i,j)\in E} \Omega_{ij} S_{ij} + \sum_{i=j} \Omega_{ij} S_{ij} \\ \implies l(\Omega) &= ce^{-\frac{n}{2}\sum_{i=1}^p \sum_{j=1}^p \Omega_{ij} S_{ij} + \frac{n}{2}\log|\Omega|} \end{aligned}$$

One choice of prior for Ω could be:

$$\tilde{\pi}_{n_0,\Lambda}(\Omega) = e^{\frac{n_0}{2}tr(\Omega\Lambda)+\frac{n_0}{2}\log|\Omega|}$$

where $\Omega \in \mathbb{P}_G, \Lambda \in \mathbb{P}^+, n_0 > 0$ not necessarily an integer. This is the kernel of the Wishart distribution. We can generalize this to the DY class of prior densities for Ω called G-Wishart with parameters $U_{p\times p} \in \mathbb{P}^+, \delta > 0$. It's density proportional is to:

$$\mathcal{GW}_p(\delta, U) \propto e^{-\frac{1}{2}tr(\Omega U)+\frac{\delta}{2}\log|\Omega|}$$

. In such a case, the posterior of $\Omega|X^1, ..., X^n \sim \mathcal{G}(n+\delta, S+U)$. Letac-Massam(2007) extends the G-Wishart priors for decomposable graphs. Note that if $K(\Omega) = \log|\Omega|$, then $\nabla K(\Omega) = \frac{1}{|\Omega|} \times |\Omega| \times (\Omega^{-1})^T = \Omega^{-1} = \Sigma$. Thus by Lemma 11, $\mathbb{E}[\nabla K(\Omega)|X^1, ...X^n] = \mathbb{E}[\Sigma|X^1, ...X^n] = \frac{n_0 x_0 + n\bar{X}}{n_0+n}$.

6.2.8 Sampling from the G-Wishart if G is decomposable

Suppose G is decomposable. Thus there is at least one perfect vertex elimination scheme. Further suppose that the vertices have been ordered according to this scheme. Then there exists a unique modified cholesky decomposition, $\Omega = LDL^T$, which is a bijection from $\mathbb{P}_G \mapsto \mathcal{L}_G \times \mathcal{D}$. Given that

$$\pi_{\delta,U}(\Omega) \propto e^{-\frac{1}{2}tr(\Omega U)+\frac{\delta}{2}\log|\Omega|}$$

we want to find $\pi_{\delta,U}(L, D) = \pi_{\delta,U}(\Omega(L, D))|\frac{\partial\Omega}{\partial(L,D)}|$. Let us consider $p = 3$ to see what's going on.

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix}$$

where $\Omega_{ij} = \Omega_{ji}$ and

$$\begin{aligned} LDL^T &= \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{pmatrix} \begin{pmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} d_{11} & d_{11}l_{21} & d_{11}l_{31} \\ 0 & d_{22} & d_{22}l_{32} \\ 0 & 0 & d_{33} \end{pmatrix} \\ &= \begin{pmatrix} d_{11} & d_{11}l_{21} & d_{11}l_{31} \\ d_{11}l_{21} & d_{22}+d_{11}l_{21}^2 & d_{11}l_{31}l_{21}+d_{22}l_{32} \\ d_{11}l_{31} & d_{11}l_{31}l_{21}+d_{22}l_{32} & d_{11}l_{31}^2+d_{22}l_{32}^2+d_{33} \end{pmatrix} \end{aligned}$$

Thus the Jacobian Matrix would look like

	Ω_{11}	Ω_{21}	Ω_{22}	Ω_{31}	Ω_{32}	Ω_{33}
d_{11}	1	l_{21}	l_{21}^2	l_{31}	$l_{31}l_{21}$	l_{31}^2
l_{21}	0	d_{11}	$2d_{11}l_{21}$	0	$l_{31}d_{11}$	0
d_{22}	0	0	1	0	l_{32}	l_{32}^2
l_{31}	0	0	0	d_{11}	$d_{11}l_{21}$	$2d_{11}l_{31}$
l_{32}	0	0	0	0	d_{22}	$2d_{22}l_{32}$
d_{33}	0	0	0	0	0	1

It is clear that the Jacobian is upper triangular and then determinant can be obtained simply by multiplying the diagonal entries together. We can also deduce that $|J| = J_{1,1}...J_{6,6} = d_{11}^2 d_{22}$. Generalizing to the case with p variables we get $|J| = J_{1,1}...J_{\frac{p(p+1)}{2},\frac{p(p+1)}{2}} = d_{11}^{n_1}d_{22}^{n_2}...d_{p-1,p-1}^{n_{p-1}}$ where $n_{p-1} = |\{i|i > j, (i,j) \in E\}|$. This is because $n_p = 0$. Now recall that a complete graph has all possible edges. i.e. a graph with p variables has $\binom{p}{2}$ edges. As the graph is decomposable note that this implies that $i > j \implies L_{ij} \neq 0$. This is the case we had in our example with $p = 3$. In such a case, $n_j = p - j$. Thus the density on $\mathcal{L}_G \times \mathcal{D}$ induced by $\pi_{\delta,U}$ is:

$$\pi_{\delta,U}(L,D) \propto e^{-\frac{1}{2}tr(LDL^TU)+\frac{\delta}{2}\log|LDL^T|}\prod_{j=1}^{p-1} d_{jj}^{n_j}.$$

As

$$\begin{aligned}\log|LDL^T| &= \log|L||D||L^T| \\ &= \log|L| + \log|D| + \log|L^T| \\ &= \log|L| + \log|D| + \log|L| \\ &= \log|D| \qquad \text{as } |L| = 1 \qquad\qquad = \prod_{i=1}^{p} d_{jj}\end{aligned}$$

and

$$\begin{aligned}
tr(LDL^TU) &= tr(DL^TUL) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}(L^TUL)_{ij} \\
&= \sum_{i=1}^{p} d_{ii}(L^TUL)_{ii} \qquad \text{as } i \neq j \implies d_{ij} = 0 \\
&= \sum_{i=1}^{p} d_{ii}(L_{.i})^T U(L_{.i}) \qquad L_{.i} \text{ is the } i\text{-th column of L} \\
&= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii} \begin{pmatrix} 1 & x_i^T \end{pmatrix} \begin{pmatrix} U_{ii} & (U_{.i}^{>})^T \\ (U_{.i}^{>}) & U^{>i} \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\
&= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii} \begin{pmatrix} 1 & x_i^T \end{pmatrix} \begin{pmatrix} U_{ii} + (U_{.i}^{>})^T x_i \\ (U_{.i}^{>}) + U^{>i} x_i \end{pmatrix} \\
&= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii}(U_{ii} + (U_{.i}^{>})^T x_i + x_i^T (U_{.i}^{>}) + x_i^T U^{>i} x_i) \\
&= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii}(U_{ii} + 2x_i^T (U_{.i}^{>}) + x_i^T U^{>i} x_i)
\end{aligned}$$

Now note that

$$\begin{aligned}
&(x_i + (U^{>i})^{-1}(U_{.i}^{>}))^T U^{>i}(x_i + (U^{>i})^{-1}(U_{.i}^{>})) \\
=&(x_i^T U^{>i} + (U_{.i}^{>})^T U^{>i}(U^{>i})^{-1})(x_i + (U^{>i})^{-1}(U_{.i}^{>})) \\
=&(x_i^T U^{>i} + (U_{.i}^{>})^T)(x_i + (U^{>i})^{-1}(U_{.i}^{>})) \\
=&x_i^T U^{>i} x_i + (U_{.i}^{>} x_i) + x_i^T U^{>i}(U^{>i})^{-1}(U_{.i}^{>}) + (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>}) \\
=&x_i^T U^{>i} x_i + U_{.i}^{>} x_i + x_i^T U^{>i} + (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>}) \\
=&x_i^T U^{>i} x_i + 2U_{.i}^{>} x_i + (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>})
\end{aligned}$$

Therefore,

$$\begin{aligned}
tr(LDL^TU) &= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii}(U_{ii} + 2x_i^T(U_{.i}^{>}) + x_i^TU^{>i}x_i) \\
&= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii}(U_{ii} + 2x_i^T(U_{.i}^{>}) + x_i^TU^{>i}x_i \\
&\quad + (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>}) - (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>})) \\
&= d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii}((x_i + (U^{>i})^{-1}(U_{.i}^{>}))^TU^{>i}(x_i + (U^{>i})^{-1}(U_{.i}^{>})) \\
&\quad + (U_{ii} - (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>})))
\end{aligned}$$

Let $c_i = (U_{ii} - (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>}))$ and $e_i = (U^{>i})^{-1}(U_{.i}^{>})$. As U is positive definite, $c_i = (U_{ii} - (U_{.i}^{>})(U^{>i})^{-1}(U_{.i}^{>})) > 0$. Thus leaving out constants

$$\begin{aligned}
\log \pi_{\delta,U}(L,D) = &\underbrace{-\frac{1}{2}tr(LDL^TU)}_{-\frac{1}{2}(d_{pp}U_{pp}+\sum_{i=1}^{p-1} d_{ii}((x_i+e_i)^TU^{>i}(x_i+e_i)+c_i))} \\
&+ \frac{\delta}{2}\log\underbrace{|LDL^T|}_{\prod_{i=1}^{p} d_{ii}} + \log\prod_{j=1}^{p-1} d_{jj}^{n_j}. \\
&= -\frac{1}{2}(d_{pp}U_{pp} + \sum_{i=1}^{p-1} d_{ii}((x_i+e_i)^TU^{>i}(x_i+e_i) + c_i) \\
&+ \frac{\delta}{2}\log\prod_{i=1}^{p} d_{ii} + \log\prod_{j=1}^{p-1} d_{jj}^{n_j} \\
&= -\frac{1}{2}d_{pp}U_{pp} - \frac{1}{2}\sum_{i=1}^{p-1} d_{ii}(x_i+e_i)^TU^{>i}(x_i+e_i) - \frac{1}{2}\sum_{i=1}^{p-1} d_{ii}c_i \\
&+ \log d_{pp}^{\frac{\delta}{2}} + \sum_{i=1}^{p-1}\log d_{ii}^{\frac{\delta}{2}} + \sum_{j=1}^{p-1}\log d_{jj}^{n_j} \\
&= \log d_{pp}^{\frac{\delta}{2}} - \frac{1}{2}d_{pp}U_{pp} \\
&\sum_{i=1}^{p-1} -\frac{1}{2}d_{ii}(x_i+e_i)^TU^{>i}(x_i+e_i)\sum_{i=1}^{p-1} -\frac{1}{2}d_{ii}c_i + \sum_{i=1}^{p-1}\log d_{ii}^{\frac{\delta}{2}+n_i}
\end{aligned}$$

Thus

$$\pi_{\delta,U}(L,D) = d_{pp}^{\frac{\delta}{2}} e^{-\frac{1}{2}d_{pp}U_{pp}} \prod_{i=1}^{p-1} d_{ii}^{\frac{\delta}{2}+n_i} e^{-\frac{1}{2}d_{ii}c_i} e^{-\frac{1}{2}d_{ii}(x_i+e_i)^T U^{>i}(x_i+e_i)},$$

which implies that $\{x_i, d_{ii}\}_{i=1}^{p}$ are independent. Recall that the density of a $Gamma(\alpha, \beta)$ distribution is

$$f(x) \propto x^{\alpha-1}e^{-\beta x}$$

Thus $d_{pp} \sim Gamma(\frac{\delta}{2}+1, \frac{U_{pp}}{2})$. Now consider the kernel of the density of $x_i|d_{ii}$ which is $e^{-\frac{1}{2}d_{ii}(x_i+e_i)^T U^{>i}(x_i+e_i)}$. Clearly this resembles a normal density with mean $e_i = (U^{>i})^{-1}(U^{>}_{.i})$ and variance $\frac{(U^{>i})^{-1}}{d_{ii}}$. Now using the simple formula that $\pi(x_i, d_{ii}) = \pi_{x_i|d_{ii}}(x_i)\pi_{d_{ii}}(d_{ii})$ we can derive that for $i = 1, ...p - 1, d_{ii} \sim Gamma(\frac{\delta}{2}+n_i+1, \frac{1}{2c_i})$. We do this by considering only the part we haven't looked at as yet:

$$d_{ii}^{\frac{\delta}{2}+n_i} e^{-\frac{1}{2}d_{ii}c_i}$$

This clearly resembles a $Gamma(\frac{\delta}{2}+n_i+1, \frac{1}{2c_i})$ with $c_i = (U_{ii} - (U^{>}_{.i})(U^{>i})^{-1}(U^{>}_{.i}))$.

Thus if G is decomposable and $\Omega \in \mathbb{P}_G$, then a sample, $\hat{\Omega}$, from the G-Wishart distribution with parameters $\delta > 0$ and $U \in \mathbb{P}^+$ is $\hat{\Omega} = \hat{L}\hat{D}\hat{L}^T$, where the i-th column of L,

$$L_{.i} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ x_i \end{pmatrix}$$

and $D = diag(d_{11}, ..., d_{pp})$ and x_i and d_{ii} are generated as described above.

6.2.9 Block Gibbs-Sampling for G-Wishart if G is not decomposable

This is due to Piccioni (2000) from the Scandinavian Journal of Statistics. Recall that the density of Ω is:

$$\pi_{\delta,U}(\Omega) \propto e^{-\frac{1}{2}tr(\Omega U)+\frac{\delta}{2}\log|\Omega|}, \Omega \in \mathbb{P}_G$$

where G is not necessarily decomposable. Let C be any clique of $G = (V, E)$ and let

$$\Omega = \begin{pmatrix} \Omega_{CC} & \Omega_{C\bar{C}} \\ \Omega_{\bar{C}C} & \Omega_{\bar{C}\bar{C}} \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} U_{CC} & U_{C\bar{C}} \\ U_{\bar{C}C} & U_{\bar{C}\bar{C}} \end{pmatrix}$$

We are interested in finding the conditional density of $\Omega_{CC}|(\Omega_{C\bar{C}}, \Omega_{\bar{C}\bar{C}})$. Thus only keeping terms containing Ω_{CC} we get that:

$$\begin{aligned} \pi_{\delta,U}(\Omega) &\propto e^{-\frac{1}{2}tr(\Omega U)+\frac{\delta}{2}\log|\Omega|} \\ &= e^{-\frac{1}{2}tr(\Omega_{CC}U_{CC})+\frac{\delta}{2}\log|\Omega_{CC}-\Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C}|} \\ &\times \text{ function of } (\Omega_{C\bar{C}}, \Omega_{\bar{C}\bar{C}}) \end{aligned}$$

Theorem 22. *Let*

$$\Omega = \begin{pmatrix} \Omega_{CC} & \Omega_{C\bar{C}} \\ \Omega_{\bar{C}C} & \Omega_{\bar{C}\bar{C}} \end{pmatrix}$$

Then Ω is positive definite if and only if

1. *$\Omega_{\bar{C}\bar{C}}$ is positive definite*
2. *$\Omega_{CC} - \Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C}$ is positive definite*

In such a case $|\Omega| = |\Omega_{\bar{C}\bar{C}}||\Omega_{CC} - \Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C}|$.

Now the parameter space for $\Omega_{CC}|(\Omega_{C\bar{C}}, \Omega_{\bar{C}\bar{C}})$ is

$$\{\Omega_{CC} : \Omega_{CC} - \Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C} \text{ is positive definite}\}$$

. Define $K_C \coloneqq \Omega_{CC} - \Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C}$. Then the conditional density of K_C

$$\pi_{K_C|(\Omega_{C\bar{C}},\Omega_{\bar{C}\bar{C}})} \propto e^{-\frac{1}{2}tr(K_C U_{CC})+\frac{\delta}{2}\log|K_C|}, K_C \in \mathbb{P}^+.$$

which we recognize as the kernel of the G-Wishart Distribution with parameters $\delta > 0$ and $U_{CC}^{-1} \in \mathbb{P}^+$. Thus $K_C|(\Omega_{C\bar{C}}, \Omega_{\bar{C}\bar{C}}) \sim \mathcal{GW}(\delta, U_{CC})$. Note that a maximal clique is one where all the vertices have an edge between them and hence the clique is a decomposable graph. Thus we can sample from this distribution using the methods discussed in the previous section to get K_{CC} and then let $\Omega_{CC} = K_C + \Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C}$. Now suppose that $C_1, ..., C_k$ be the collection of maximal cliques of G. Then the block Gibbs sampling algorithm for $\Omega \in \mathbb{P}_G$ where G is not necessarily decomposable is as follows:

1. Start with initial value $\Omega^{(0)}$. Set $r = 0$ and $\Omega^{(r,0)} = \Omega^{(0)}$.
2. Repeat for $i = 1, ..., k$ Obtain $\Omega^{(r,i)}$ by sampling from the conditional distribution of $\Omega_{C_iC_i}|(\Omega_{C\bar{C}} = \Omega^{(r,i-1)}_{C\bar{C}}, \Omega_{\bar{C}\bar{C}} = \Omega^{(r,i-1)}_{\bar{C}\bar{C}})$.
3. If convergence criterion is met, then stop and accept $\Omega^{(r,k)}$ as a sample. Otherwise set $\Omega^{(r+1,0)} = \Omega^{(r,k)}$ and return to Step 2.

Note that while in standard gibbs-sampling we consider a partition of the sampling space, in this case $C_1, ..., C_k$ is not a partition but a cover of the space. However, Piccioni(2000) provides justification for the convergence of the Markov chain produced in this manner. The reason we use this method is that it is easy to find the distribution of $\Omega_{C_iC_i}|(\Omega_{C\bar{C}}, \Omega_{\bar{C}\bar{C}})\forall i = 1, ..., k$ and also to generate from it.

Lenoski(2013) suggest another method for sampling from a G-Wishart distribution when G is not decomposable. In fact this is an exact sampling method.

1. Generate $\Lambda \sim \mathcal{W}_p(\delta, U)$.
2. Let $\hat{\Omega} = \arg\min_{\Omega \in \mathbb{P}_G}\{tr(\Omega\Lambda) - \log|\Omega|\}$.

Then $\hat{\Omega} \sim \mathcal{GW}_p(\delta, U)$.

6.2.10 Other Sampling Mechanisms for the G-Wishart

Let $\Omega = \Phi^T\Phi$ be the Cholesky decomposition of Ω, where Φ is an upper triangular matrix and $\Phi_{ii} > 0 \forall i = 1, ..., p$. If G is decomposable then $(i, j) \notin E \implies \Phi_{ij} = 0$. However we are interested in the case when G is not decomposable. Then,

$$\pi_{\delta,U}(\Omega) \propto e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|}, \Omega \in \mathbb{P}_G$$

7 Covariance Graph Models

We want to estimate Σ where $\mathbf{Y}^1, ..., \mathbf{Y}^n \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0}, \Sigma)$. We are especially interested in the case when $n < p$. Recall that Σ has be be positive definite. As $\mathbf{Y}_i\mathbf{Y}_i^T$ is a $p \times p$ matrix of rank 1, $S = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i\mathbf{Y}_i^T$ can be full rank and hence positive definite if $n \geq p$. Intuitively, if $\mathbf{Y}^1 \perp\!\!\!\perp \mathbf{Y}^2$, then $Cov(\mathbf{Y}^1, \mathbf{Y}^2) = \mathbb{E}[\mathbf{Y}^1(\mathbf{Y}^2)^T] = 0 \implies \forall \epsilon > 0, P((\mathbf{Y}^2)^T\mathbf{Y}^1 > \epsilon) \leq \frac{\mathbb{E}[(\mathbf{Y}^2)^T\mathbf{Y}^1]}{\epsilon} = \frac{\mathbb{E}[tr((\mathbf{Y}^2)^T\mathbf{Y}^1)]}{\epsilon} = maybeusematrixasAandBneedstobesquare? = 0$. Thus, with probability 1, we can say that $\mathbf{Y}^1$ and $\mathbf{Y}^2$ are linearly independent. Thus if $n \geq p$, we can use S as an estimate of Σ. However, if $n < p$, then S has rank less than p and thus it cannot be positive definite. Thus we cannot use S as an estimate of Σ. Estimating Σ consists of two steps. In our experience it is always better to divide a task up in to smaller, simpler tasks if possible.

1. Identify elements of Σ that we want to set to 0. This means find (i, j) such that $(i, j) \notin E$, where E is the edge set for the graph $G = (V, E)$ that encodes the non-zero elements of Σ. The simplest way to do this is through thresholding, which essentially involves setting the off-diagonal elements of S to 0 if they are below a certain threshold, thr. Thus $(S^{thr})_{ij} = \begin{cases} S_{ij}, & \text{if } S_{ij} > thr \\ 0, & \text{if } S_{ij} \leq thr \end{cases}$.

2. Now use the data ({(i,j): $S_{ij} < thr$}) to form constraints and $S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i\mathbf{Y}_i^T$ to find our positive definite estimate, $\hat{\Sigma}$, of Σ.

7.1 Graphical Lasso

Let

$$Q_{GL}(\Omega) = tr(\Omega S) - \log|\Omega| + \lambda \sum_{1 \leq i < j \leq p} |\Omega_{ij}|, \quad \Omega \in \mathbb{P}_G$$

denote the objective function for graphical lasso. We want to find $\hat{\Omega}$ such that $Q_{GL}(\hat{\Omega}) = \arg\min_{\Omega \in \mathbb{P}_G} Q_{GL}(\Omega)$. Fix $i \in \{1, ..., p\}$. Let

$$\begin{aligned} \Omega_{-i,i} &= (\Omega_{ji})_{j \neq i} \\ \Omega_{-i,-i} &= (\Omega_{kl})_{k,l \neq i} \\ \gamma_i &= \Omega_{i,i} - \Omega_{i,-i}\Omega_{-i,-i}^{-1}\Omega_{-i,i} \end{aligned}$$

Then

$$\Omega \mapsto (\Omega_{-i,i}, \Omega_{-i,-i}, \gamma_i)$$

is a bijection. An ℓ_1 penalty, as we have here, imposes both sparsity and shrinkage as opposed to a ridge penalty which imposes only shrinkage. Also, as $|\Omega| = |\Omega_{-i,-i}||\gamma_i|$ by Equation 1,

$$\begin{aligned}
Q_{GL}(\Omega) &= tr(\Omega S) - \log|\Omega| + \lambda \sum_{1\le i<j\le p} |\Omega_{ij}| \\
&= tr\Big(\begin{pmatrix} \Omega_{i,i} & \Omega_{i,-i} \\ \Omega_{-i,i} & \Omega_{-i,-i} \end{pmatrix} \begin{pmatrix} S_{i,i} & S_{i,-i} \\ S_{-i,i} & S_{-i,-i} \end{pmatrix} \Big) - \log|\Omega_{-i,-i}||\gamma_i| \\
&\quad + \lambda \sum_{1\le i<j\le p} |\Omega_{ij}| \\
&= tr\Big(\begin{pmatrix} \Omega_{i,i} & \Omega_{i,-i} \\ \Omega_{-i,i} & \Omega_{-i,-i} \end{pmatrix} \begin{pmatrix} S_{i,i} & S_{i,-i} \\ S_{-i,i} & S_{-i,-i} \end{pmatrix} \Big) - \log|\Omega_{-i,-i}| - \log|\gamma_i| \\
&\quad + \lambda\|\Omega_{-i,i}\|_1 + \text{ other terms depending on } \Omega_{-i,-i} \\
&= tr(\Omega_{i,i}S_{i,i} + \Omega_{i,-i}S_{-i,i} + \Omega_{-i,i}S_{i,-i} + \Omega_{-i,-i}S_{-i,-i}) - \log|\gamma_i| \\
&\quad + \lambda\|\Omega_{-i,i}\|_1 + \text{ other terms depending on } \Omega_{-i,-i} \\
&= tr(\gamma_i S_{i,i} + \Omega_{i,-i}\Omega^{-1}_{-i,-i}\Omega_{-i,i}S_{i,i} + \Omega_{i,-i}S_{-i,i} + \Omega_{-i,i}S_{i,-i}) \\
&\quad + \lambda\|\Omega_{-i,i}\|_1 + \text{ other terms depending on } \Omega_{-i,-i}, \gamma_i \\
&= \Omega_{i,-i}\Omega^{-1}_{-i,-i}\Omega_{-i,i}S_{i,i} + 2\Omega_{i,-i}S_{-i,i} \\
&\quad + \lambda\|\Omega_{-i,i}\|_1 + \text{ other terms depending on } \Omega_{-i,-i}, \gamma_i
\end{aligned}$$

as $tr(\Omega_{i,-i}S_{-i,i}) = \Omega_{i,-i}S_{-i,i} = \Omega_{-i,i}S_{i,-i} \in \mathbb{R}$ and $tr(\Omega_{i,-i}\Omega^{-1}_{-i,-i}\Omega_{-i,i}S_{i,i}) = \Omega_{i,-i}\Omega^{-1}_{-i,-i}\Omega_{-i,i}S_{i,i} \in \mathbb{R}$. Thus, as $S_{i,i}$ is a constant

$$\implies Q_{GL}(\Omega|\Omega_{-i,i}, \gamma_i) = \Omega^T_{-i,i}(S_{i,i}\Omega^{-1}_{-i,-i})\Omega_{-i,i} + 2\Omega_{-i,i}S_{i,-i} + \lambda\|\Omega_{-i,i}\|_1$$

which implies that keeping γ_i and $\Omega_{-i,-i}$ fixed and then minimizing Q_{GL} with respect to $\Omega_{i,-i}$ is equivalent to a regression lasso problem. Similarly we can show that

$$\implies Q_{GL}(\gamma_i|\Omega_{-i,i}, \Omega_{-i,-i}) = \gamma_i S_{i,i} - \log\gamma_i$$

which is minimized at $\hat{\gamma}_i = \frac{1}{S_{ii}}$. We can use coordinate-wise minimization to update γ_i and $\Omega_{-i,i}$ for $i = 1, ..., p$. Convergence is guaranteed by convexity of the objective function. In general, the algorithm is $O(p^4)$ as each iteration involves inverting a $(p-1)\times(p-1)$ matrix, which is $O(p-1)^3 \approx O(p^3)$ and there are p iterations in total.

7.2 SPACE algorithm

This is due to Peng et.al. (2009) in JASA. The main idea here is to change the objective function by replacing the likelihood with the psuedolikelihood, which is the product of full conditionals. Note that if $Y = (Y_1, ...Y_p) \sim \mathcal{N}_p(0, \Sigma = \Omega^{-1})$, then

$$Y_1|Y_{-1} \sim \mathcal{N}(\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(Y_{-1}), \Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})$$

Now note that $\Omega_{11} = \frac{1}{\Sigma_{1,1}-\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}}$. For the next part we need the following result:

Lemma 12. $\Sigma_{i,-i}\Sigma_{-i,-i} = -(\Omega_{ii})^{-1}\Omega_{i,-i}$.

This tells us that

$$\begin{aligned}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(Y_{-1}) &= -(\Omega_{11})^{-1}\Omega_{1,-1}(Y_{-1}) \\ &= -\frac{1}{\Omega_{11}}(\Omega_{12}, ..., \Omega_{1p})\begin{pmatrix} Y_2 \\ \vdots \\ Y_p \end{pmatrix} \\ &= -\sum_{j \neq 1} \frac{\Omega_{j1}}{\Omega_{11}} Y_j\end{aligned}$$

In general,

$$Y_i|Y_{-i} \sim \mathcal{N}(-\sum_{j \neq i} \frac{\Omega_{ji}}{\Omega_{ii}} Y_j, \frac{1}{\Omega_{ii}}).$$

Thus the psuedolikelihood is

$$\begin{aligned}p(\Omega) &= \prod_{k=1}^{n}\prod_{i=1}^{p} f_{Y_i|Y_{-i}}(Y_i^k) \\ &= \prod_{k=1}^{n}\prod_{i=1}^{p} \frac{\sqrt{\Omega_{ii}}}{\sqrt{2\pi}} e^{-\frac{\Omega_{ii}}{2}(Y_i^k - (-\sum_{j \neq i}\frac{\Omega_{ji}}{\Omega_{ii}}Y_j^k))^2}\end{aligned}$$

and the negative log of the psuedolikelihood is

$$\begin{aligned}
-\log p(\Omega) &= \sum_{k=1}^{n}\sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}(Y_i^k - (-\sum_{j\neq i}\frac{\Omega_{ji}}{\Omega_{ii}}Y_j^k))^2 - \frac{1}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\sum_{k=1}^{n}(Y_i^k - (-\sum_{j\neq i}\frac{\Omega_{ji}}{\Omega_{ii}}Y_j^k))^2 - \frac{n}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\sum_{k=1}^{n}(\sum_{j=1}^{p}\frac{\Omega_{ji}}{\Omega_{ii}}Y_j^k))^2 - \frac{n}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\frac{1}{\Omega_{ii}^2}\sum_{k=1}^{n}(\sum_{j=1}^{p}\Omega_{ji}Y_j^k))^2 - \frac{n}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\frac{1}{\Omega_{ii}^2}\sum_{k=1}^{n}(\Omega_{1i}Y_1^k + ... + \Omega_{pi}Y_p^k)^2 - \frac{n}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\frac{1}{\Omega_{ii}^2}\sum_{k=1}^{n}((Y^k)^T\Omega_{.i})^2 - \frac{n}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\frac{1}{\Omega_{ii}^2}\sum_{k=1}^{n}((Y^k)^T\Omega_{.i})^T((Y^k)^T\Omega_{.i}) - \frac{n}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\frac{1}{\Omega_{ii}^2}\Omega_{.i}^T nS\Omega_{.i} - \frac{n}{2}\log\Omega_{ii}\}
\end{aligned}$$

Thus for one observation the negative log psuedolikelihood is

$$= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\frac{1}{\Omega_{ii}^2}\Omega_{.i}^T S\Omega_{.i} - \frac{1}{2}\log\Omega_{ii}\}$$

Note that we can also write

$$\begin{aligned}
-\log p(\Omega) &= \sum_{k=1}^{n}\sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}(Y_i^k - (-\sum_{j\neq i}\frac{\Omega_{ji}}{\Omega_{ii}}Y_j^k))^2 - \frac{1}{2}\log\Omega_{ii}\} \\
&= \sum_{i=1}^{p}\{\frac{\Omega_{ii}}{2}\|(Y_i^k - (-\sum_{j\neq i}\frac{\Omega_{ji}}{\Omega_{ii}}Y_j^k))\|_2^2 - \frac{n}{2}\log\Omega_{ii}\}
\end{aligned}$$

If the main objective is to estimate the sparsity pattern, then the restriction $\Omega \in \mathbb{P}_G$ or $\Omega \in \mathbb{P}^+$ is not needed. Define

$$\beta_{ij} = -\frac{\Omega_{ij}}{\Omega_{ii}}$$
$$\beta = (\beta_{ij})_{1 \leq i < j \leq p}$$

and the partial correlation coefficient

$$\rho_{ij} = \frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}$$
$$\rho = (\rho_{ij})_{1 \leq i < j \leq p}$$

Then

$$\Omega \mapsto (\beta, (\Omega_{ii})_{i=1}^p) \mapsto (\rho, (\Omega_{ii})_{i=1}^p)$$

are all bijections, which means that sparsity in Ω is equivalent to sparsity in ρ and β. Thus if $Y^1, ..., Y^n \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma = \Omega^{-1})$, let $Y_i = (Y_i^1, ..., Y_i^n)$ is the i-th component of all the observations put into a $n \times 1$ vector. Then the objective function of the SPACE algorithm is:

$$Q_{SPACE}(\rho, (\Omega_{ii})_{i=1}^p) = \sum_{i=1}^p \frac{\Omega_{ii}}{2} \|Y_i - (-\sum_{j \neq i} \beta_{ij} Y_j)\|_2^2 - \frac{n}{2} \sum_{i=1}^p \log(\Omega_{ii}) + \lambda \sum_{1 \leq i < j \leq p} |\rho_{ij}|$$

We can fo the minimization of the objective function in two steps. First fix Ω and estimate β by adaptive lasso regressions. Then fix β, which has a closed form solution. One thing to note is that postive definiteness is lost in this approach, but that is of little concern as the main interest here is the sparsity pattern. The objective function $Q_{SPACE}(\rho, (\Omega_{ii})_{i=1}^p)$ is not jointly convex, but is bi-convex, which means that it is convex in each part keeping the other part constant. As a result, we can construct some examples where the SPACE algorithm does not converge.

```
SPACE pseudocode
Input: Standardize data to have mean zero and standard deviation one
Input: Fix maximum number of iterations: $r_{max}$
Input: Fix initial estimate: $\hat{\Omega}_{ii}^{(0)} = \frac{1}{S_{ii}}$ as suggested
Input: Choose weights: $w_i (w_i = \Omega_{ii}$ or $w_i = 1)$
Set $r \leftarrow 1$
repeat
## update partial correlations
    Update $\hat{\rho}^{(r)}$ by minimizing (with current estimates $\{\hat{\Omega}_{ii}^{(r-1)}\}_{i=1}^{p}$)

  $\frac{1}{2}(\sum_{i=1}^{p} w_i \|Y_i - (-\sum_{j \neq i} \rho_{ij} \sqrt{\frac{\Omega_{jj}^{(r-1)}}{\Omega_{ii}^{(r-1)}}} Y_j)\|_2^2) + \lambda \sum_{1 \leq i < j \leq p} |\rho_{ij}|$

## update conditional variance
    Update $\{\Omega_{ii}^{(r)}\}_{i=1}^{p}$ by computing (with fixed estimates $\{\hat{\rho}_{ij}^{(r-1)}\}$
    and $\{\hat{\Omega}_{ii}^{(r-1)}\}_{i=1}^{p}$)

  $\frac{1}{\hat{\Omega}_{ii}^{r}} = \frac{1}{n} \|Y_i - (-\sum_{j \neq i} \hat{\rho}_{ij}^{(r-1)} \sqrt{\frac{\Omega_{jj}^{(r-1)}}{\Omega_{ii}^{(r-1)}}} Y_j)\|_2^2$

    for $i = 1, ..., p$
    $r \leftarrow r + 1$
    Update weights: $w_i$
until $r == r_{max}$
Return $(\hat{\rho}^{(r_{max})}, \{\hat{\Omega}_{ii}^{(r_{max})}\}_{i=1}^{p})$
```

7.3 CONCORD algorithm

This is due to Khare, Oh and Rajaratnam (2014) in JRSSB. First note that the objective function of the SPACE algorithm can be rewritten as:

$$Q_{SPACE}(\Omega) = \frac{n}{2} \sum_{i=1}^{p} \frac{w_i}{\Omega_{ii}^2} (\Omega_{.i}^T S \Omega_{.i}) - \frac{n}{2} \sum_{i=1}^{p} \log(\Omega_{ii}) + \lambda \sum_{1 \leq i < j \leq p} |\frac{\Omega_{ij}}{\sqrt{\Omega_{ii} \Omega_{jj}}}|$$

where w_i is a weight variable. Peng et. al. suggest using $w_i = 1$ or $w_i = \Omega_{ii}$, but neither choice guarantees convergence. The main idea of CONCORD is to make some adjustment to $Q_{SPACE}(\Omega)$ to make it jointly convex. The changes in $Q_{SPACE}(\Omega)$ to make it resemble -2loglikelihood plus a penalty term are listed below:

1. $w_i = \Omega_{ii}^2$
2. change $|\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}|$ with Ω_{ij}
3. multiply $\frac{1}{2}\sum_{i=1}^{p}\log(\Omega_{ii})$ by 2

Thus the objective function for the CONCORD algorithm is

$$Q_{CONCORD}(\Omega) = \frac{n}{2}\sum_{i=1}^{p}(\Omega_{.i}^T S \Omega_{.i}) - n\sum_{i=1}^{p}\log(\Omega_{ii}) + \lambda \sum_{1\leq i<j\leq p}|\Omega_{ij}|$$

If $i \neq j$, then

$$\begin{aligned} Q_{CONCORD}(\Omega_{ij}|\Omega_{-(ij)}) = &\frac{n}{2}(S_{ii}+S_{jj})\Omega_{ij}^2 + n(\sum_{k\neq i}\Omega_{ik}S_{jk} + \sum_{k\neq j}\Omega_{jk}S_{ik})\Omega_{ij} + \lambda|\Omega_{ij}| \\ &+ \text{ terms not depending on } \Omega_{ij}, \end{aligned}$$

which look like a *lasso* problem. If $i = j$, then

$$\begin{aligned} Q_{CONCORD}(\Omega_{ii}|\Omega_{-(ii)}) &= \frac{n}{2}S_{ii}\Omega_{ii}^2 + n(\sum_{k\neq i}S_{ik}\Omega_{ki})\Omega_{ii} - n\log(\Omega_{ii}) \\ &\quad + \text{ terms not depending on } \Omega_{ii} \\ \implies \frac{d}{d\Omega_{ii}}Q_{CONCORD}(\Omega_{ii}|\Omega_{-(ii)}) &= n\Omega_{ii}S_{ii} + n(\sum_{k\neq i}S_{ik}\Omega_{ki}) - \frac{n}{\Omega_{ii}} \overset{\text{set}}{=} 0. \end{aligned}$$

Thus,

$$\Omega_{ii}^2 S_{ii} + (\sum_{k\neq i}S_{ik}\Omega_{ki})\Omega_{ii} - 1 = 0,$$

which implies that

$$\Omega_{ii} = \frac{-\sum_{k\neq i}S_{ik}\Omega_{ki} \pm \sqrt{(\sum_{k\neq i}S_{ik}\Omega_{ki})^2 + 4S_{ii}}}{2S_{ii}}.$$

As $\Omega_{ii} > 0$,

$$\Omega_{ii} = \frac{-\sum_{k\neq i} S_{ik}\Omega_{ki} + \sqrt{(\sum_{k\neq i} S_{ik}\Omega_{ki})^2 + 4S_{ii}}}{2S_{ii}}.$$

We now iterate the above 2-stage optimization for $i = 1, ..., p$ until convergence. The computational cost for each iteration is $min\{O(np^2), O(p^3)\}$ and the algorithm has consistent results in high-dimensional settings. Additionally, if $n > p$, then $Q_{CONCORD}$ is strictly convex. If $n < p$, then $Q_{CONCORD}$ is no longer strictly convex. However, convergence can still be established rigorously, even though the starting point or initial value may be a factor. Note that we can also write the objective function of CONCORD as:

$$Q_{CONCORD}(\Omega) = \frac{1}{2}\sum_{i=1}^{p} \|\Omega_{ii}Y_i + \sum_{j\neq i} \Omega_{ij}Y_j\|_2^2 - n\sum_{i=1}^{p} \log(\Omega_{ii}) + \lambda \sum_{1\leq i<j\leq p} |\Omega_{ij}|$$

We now restate the above results formally as a lemma, which is exactly the same as Lemma 4 in the paper. Let $\mathcal{A}_p$ denote the set of $p \times p$ real symmetric matrices. Let the parameter space $\mathcal{M}$ be defined as

$$\mathcal{M} := \{\Omega \in \mathcal{A} : \Omega_{ii} > 0 \forall 1 \leq\leq p\}$$

For $1 \leq i \leq j \leq p$, define $T_{ij} : \mathcal{M} \mapsto \mathcal{M}$ by

$$T_{ij}(\Omega) = \arg \min_{\tilde{\Omega}:\tilde{\Omega}_{kl}=\Omega_{kl}\forall(k,l)\neq(i,j)} Q_{CONCORD}(\tilde{\Omega}).$$

Thus for each (i, j), $T_{ij}(\Omega)$ gives the matrix where all the elements of Ω are left as is except the (i, j)-th element. The (i, j)-th element is replaced by the value that minimizes $Q_{CONCORD}(\Omega)$ with respect to Ω_{ij} holding all other variables $\Omega_{kl}, (k, l) \neq (i, j)$ constant.

Lemma 13. *The function $T_{ij}(\Omega)$ defined above can be computed in closed form. In particular for $1 \leq i \leq p$,*

$$(T_{ii}(\Omega))_{ii} = \frac{-\sum_{k\neq i} S_{ik}\Omega_{ki} + \sqrt{(\sum_{k\neq i} S_{ik}\Omega_{ki})^2 + 4S_{ii}}}{2S_{ii}}.$$

For $1 \leq i < j \leq p$

$$(T_{ij}(\Omega))_{ij} = \frac{S_{\frac{\lambda}{n}}(-(\sum_{j\neq j'} \Omega_{ij'}S_{jj'} + \sum_{i'\neq i} \Omega_{i'j}S_{ii'}))}{S_{ii} + S_{jj}}$$

where $S_\lambda(s) := sign(x)(|x| - \lambda)_+$ is the soft-thresholding operator.

```
CONCORD pseudocode
Input: Standardize data to have mean zero and standard deviation one
Input: Fix maximum number of iterations: $r_{max}$
Input: Fix initial estimate: $\hat{\Omega}_{ii}^{(0)} =$
Input: Fix convergence threshold: $\epsilon$
Set $r \leftarrow 1$
Set converged = FALSE
**repeat**
    $\hat{\Omega}^{old} = \hat{\Omega}^{current}$
## *updates to partial covariances* $\Omega_{ij}$
    for $i \leftarrow 1, ...p-1$ do
      for $j \leftarrow 1, ...p-1$ do

        $\hat{\Omega}_{ij}^{current} = (T_{ij}(\Omega^{current}))_{ij}$

      end for
    end for
## *updates to partial variances* $\Omega_{ii}$
    for $i \leftarrow 1, ...p-1$ do

        $\hat{\Omega}_{ii}^{current} = (T_{ii}(\Omega^{current}))_{ii}$

    end for
## *Convergence checking*
    if $\|\hat{\Omega}^{old} - \hat{\Omega}^{current}\|_{max} < \epsilon$ then
      converged = TRUE
    else
      $r \leftarrow r + 1$
    end if
**until** converged=TRUE or $r > r_{max}$
Return $(\hat{\Omega^{(r)}})$
```