Notes on Covariance Estimation

Contents

1 Introduction

Consider the following problem: We want to estimate $\Sigma \in \mathbb{R}^{p\times p}$ where $Y^1, ..., Y^n \overset{iid}{\sim} (0, \Sigma)$. 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 $X \sim \mathcal{N}_p(\mu, \Sigma)$ if and only if*

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

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

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

Lemma 2. *Now suppose $A \in \mathbb{R}^{m\times p}$. Also suppose $rank(A) = m$ and $b \in \mathbb{R}^m$. Then $AX + b \in \mathbb{R}^m$ and $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$. Let $X = \left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right) \sim \mathcal{N}(\left(\begin{smallmatrix} \mu_1 \\ \mu_2 \end{smallmatrix}\right), \left(\begin{smallmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{smallmatrix}\right))$. Then*

1. *$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$*

2. *$X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$*

3. *$X_2|X_1 \sim \mathcal{N}_p(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$*

Additionally,for normally distributed random vectors, $X_1 \perp\!\!\!\perp X_2$ if and if $\Sigma_{12} = 0$ and $\Sigma_{21} = 0$. In such a case, $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_{11} + \Sigma_{22})$ provided the addition makes sense.

2.1 Maximum Likelihood Estimation

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

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

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

Thus,

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

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

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).*

Proof. We want to show that $P(rank(S) \geq p) = 1$. For simplicity we assume $n = p$. The case when $n \geq p$ follows. First note that $Y^1, ..., Y^p \overset{iid}{\sim} \mathcal{N}_p(0,\Sigma) \implies \{Y^1, ..., Y^p\}$ has p linearly independent vectors which implies that $rank(Y^1, ..., Y^p) = p$ with probability 1. This is because if Y^1 is linearly dependent on Y^2, then $Y^1 = cY^2 \implies P(Y^1 - cY^2 = 0) = 1$. But this is impossible as $Y^1, Y^2 \overset{iid}{\sim} \mathcal{N}_p(0,\Sigma) \implies Y^1 - cY^2 \sim \mathcal{N}_p(0,(1+c^2)\Sigma)$. Thus, with probability 1, we can say that Y^1 and Y^2 are linearly independent. Now note that

$$\begin{aligned}
S &= \sum_{i=1}^{p} Y^i (Y^i)^T \\
&= (Y^1, ..., Y^p) \begin{pmatrix} (Y^1)^T \\ \vdots \\ (Y^p)^T \end{pmatrix} \\
&= \begin{pmatrix} (Y^1)^T \\ \vdots \\ (Y^p)^T \end{pmatrix}^T (Y^1, ..., Y^p)^T \\
&= S^T.
\end{aligned}$$

Thus $P(S = S^T) = 1$ and $rank(S) \geq rank((Y^1, ..., Y^p)) = p$ with probability 1, which implies that S is positive definite with probability 1. □

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 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 $M/D = A - BD^{-1}C$.

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. $(u_i, u_{i+1}) \in E, i = 1, ..., p.$
2. $(u_p, u_1) \in E.$

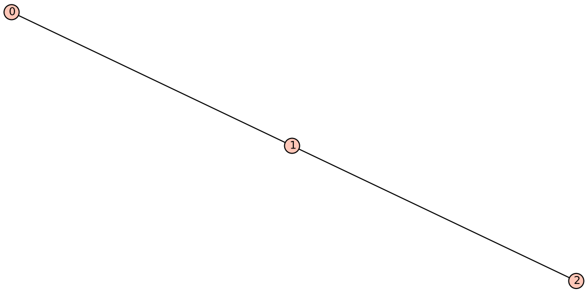


Figure 1: A graph with $V = \{0, 1, 2\}$ and $E = \{(0, 1), (1, 2)\}$.

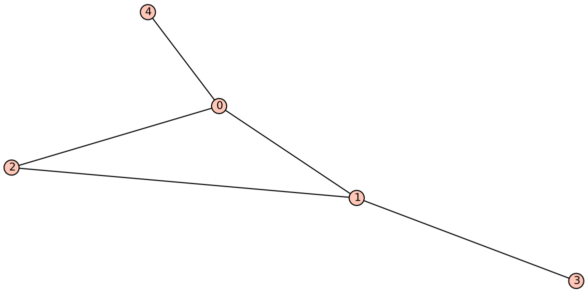


Figure 2: A graph with $V = \{0, 1, 2, 3, 4\}$ and $E = \{(0, 1), (0, 2), (0, 4), (1, 2), (1, 3)\}$.

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 Model corresponding to the Concentration Graph

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.

Model Corresponding to Graph, $G = (V, E)$**:** Let $\Omega \in \mathbb{P}_G \coloneqq \{\ A | A \in \mathbb{P}^+$ 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*

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

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.)

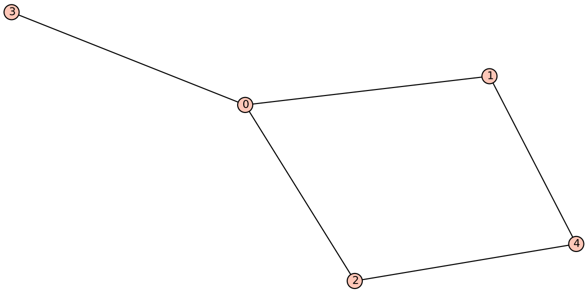


Figure 3: A graph with $V = \{0, 1, 2, 3, 4\}$ and $E = \{(0, 1), (0, 2), (0, 3), (1, 4), (2, 4)\}$.

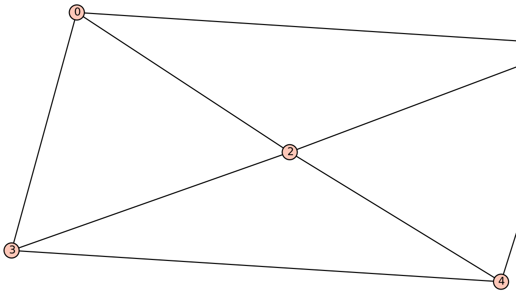


Figure 4: A graph with $V = \{0, 1, 2, 3, 4\}$ and $E = \{(0, 1), (0, 2), (0, 3), (1, 2), (1, 4), (2, 3), (3, 4)\}$.

6.3 (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.

6.4 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

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

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 $\Sigma_{\bar{C_1}\bar{C_1}}$.

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

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 Eichler (2006).
Details

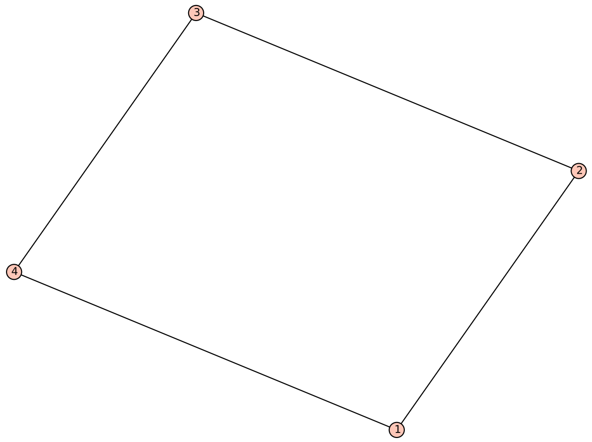
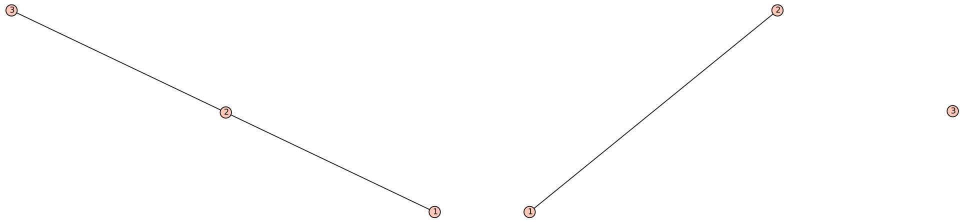


Figure 5: A graph with $V = \{1, 2, 3, 4\}$, $E = \{(4, 1), (1, 2), (2, 3), (3, 4)\}$.



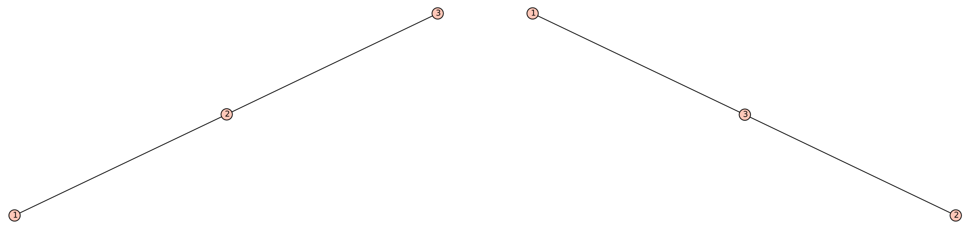
(a) An induced graph with $V_1 = \{1, 2, 3\} \subset V$, $E_1 = \{(2, 3), (1, 2)\} \subset E$.

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

Figure 6: Subgraphs

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.*

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 $(i, k) \in E_\sigma$.*



(a) Not a perfect elimination ordering of the vertices

(b) A perfect elimination ordering of the vertices

Figure 7: Perfect Elimination Orderings

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. $d_{ii} > 0 \forall i$)*

6.5 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.

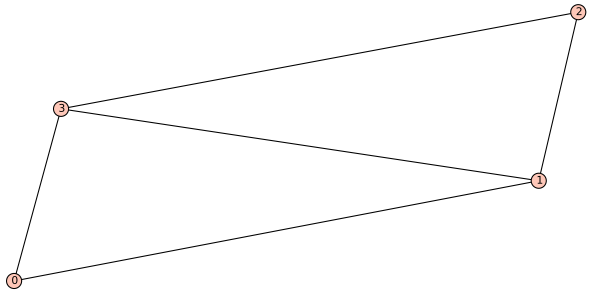


Figure 8: A graph with $V = \{0, 1, 2, 3\}$, $E = \{(0, 1), (0, 3), (1, 2), (1, 3), (2, 3)\}$.

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, $C_1, ..., C_{h-1}$.

6.6 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, i.e. $\forall i, L(\theta^*) = L\{T_i(\theta^*)\}$ As the partial maxima are unique, we also have that

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

Finally since the global maximum, $\hat{\theta}$, was uniquely determined by maximizing L over all sections, i.e. $\hat{\theta} = T_1(\theta^*) = ... = T_k(\theta^*)$, the proof is complete. □

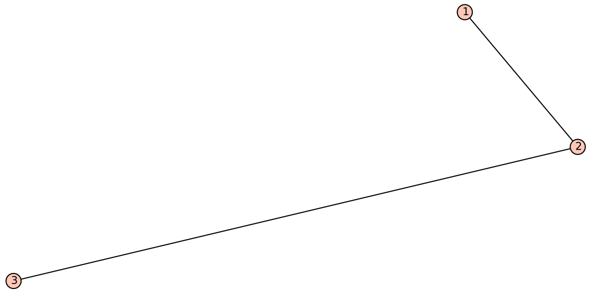


Figure 9: A graph with $V = \{1, 2, 3\}$, $E = \{(1, 2), (2, 3)\}$.

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

Let

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

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

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

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, $(D_{11}(L^T_{.1}SL_{.1}) - \log D_{11}$ and then move on to $i = 2$. In this regard note that

$$L^T_{.i}SL_{.i} = \begin{pmatrix}1 & x_i^T\end{pmatrix}\begin{pmatrix}S_{ii} & S^>_{.i}\\ S^>_{.i} & S^{>i}\end{pmatrix}\begin{pmatrix}1\\ x_i\end{pmatrix} \tag{3}$$

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,

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

6.8 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 $\tilde{\pi}_{n_0,x^0}(\theta)$ 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

$$\tilde{\pi}_{n_0,x^0}(\theta) = e^{n_0 x_0' \theta - n_0 \kappa(\theta)}.$$

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*

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

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 $nS = \sum_{i=1}^n X_iX_i^T \sim \mathcal{W}_p(n, \Sigma)$. 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 $\frac{n_0x_0+n\bar{X}}{n_0+n} = \frac{\Lambda_0+nS}{m+n-p-1}$ or directly,

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

6.9 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, $\mathbb{E}[\nabla K(\Omega)|X^1, ...X^n] = \mathbb{E}[\Sigma|X^1, ...X^n] = \frac{n_0x_0+n\bar{X}}{n_0+n}$.

6.10 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 $n_j = |\{i | i > j, (i,j) \in E\}|$. 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

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

and

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

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,

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

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

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

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 $e^{-\frac{1}{2}d_{ii}(x_i+e_i)^TU^{>i}(x_i+e_i)}$. 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.

6.11 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 $K_C := \Omega_{CC} - \Omega_{C\bar{C}}\Omega_{\bar{C}\bar{C}}^{-1}\Omega_{\bar{C}C}$. 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 $\Omega_{C_iC_i}|(\Omega_{C\bar{C}} = \Omega_{C\bar{C}}^{(r,i-1)}, \Omega_{\bar{C}\bar{C}} = \Omega_{\bar{C}\bar{C}}^{(r,i-1)})$.

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

Details

6.12 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,

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

Note that

- If G is not decomposable, then $(i,j) \in E$ does not necessarily imply that $\Phi_{ij} = 0$.
- If G is decomposable, there is a $1 - to - 1$ transformation from $\Omega \to (L, D)$ where $\Omega = LDL^T$ and

$$\Omega = \begin{cases} (\Omega_{ij}) & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

$$L = \begin{cases} 1 & \text{on the diagonal} \\ (L_{ij}) & \text{if } i > j, (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases}$$

i.e. the 0's in Ω correspond to the 0's in L. Note that this means that the 0's in Ω correspond to the 0's in Φ as well because

$$\begin{aligned}\Phi &= D^{\frac{1}{2}}L^T\\ &= \begin{pmatrix} d_{11} & 0 & \dots & 0\\ 0 & d_{22} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & d_{pp}\end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} l_{11} & l_{21} & \dots & l_{p1}\\ 0 & l_{22} & \dots & l_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & l_{pp}\end{pmatrix}\\ &= \begin{pmatrix} d_{11}^{\frac{1}{2}}l_{11} & d_{11}^{\frac{1}{2}}l_{21} & \dots & d_{11}^{\frac{1}{2}}l_{p1}\\ 0 & d_{22}^{\frac{1}{2}}l_{22} & \dots & d_{22}^{\frac{1}{2}}l_{p2}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & d_{pp}^{\frac{1}{2}}l_{pp}\end{pmatrix}\end{aligned}$$

. Thus we have that $i > j, (i., j) \notin E \implies l_{ij} = 0 \implies \Phi_{ji} = d_{jj}^{\frac{1}{2}} l_{ij} = 0$

Lemma 12. *Suppose G is not decomposable. Then the following hold:*

1. *If $(k, l) \notin E, \Phi_{kl}$ can be shown to be a function of $\{\Phi_{ij}\}_{i \leq j, (i,j) \in E}$. Thus when $(i, j) \in E, \Phi_{ij}$ are called the independent entries and when $(k, l) \notin E, \Phi_{kl}$ are called the dependent entries.*

2. *The function which maps $\{\Omega_{ij}\}_{i \leq j, (i,j) \in E}$ to the indenedent entries of $\Phi = \{\Phi_{ij}\}_{i \leq j, (i,j) \in E}$ is a bijection.*

3. *The determinant of the Jacobian matrix, which informally we can denote as $|\frac{\partial \Omega}{\partial \Phi}|$*

$$|\frac{\partial \Omega}{\partial \Phi}| = \prod_{i=1}^{p} v_i \Phi_{ii}^{v_i + 1}$$

where $v_i = |\{j : j > i, (i, j) \in E\}|$.

Proof. Suppose

$$\begin{aligned}
\Omega &= \begin{pmatrix} \Omega_{11} & \Omega_{12} & \dots & \Omega_{1p} \\ \Omega_{12} & \Omega_{22} & \dots & \Omega_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_{1p} & \Omega_{2p} & \dots & \Omega_{pp} \end{pmatrix} \\
&= \begin{pmatrix} \phi_{11} & 0 & \dots & 0 \\ \phi_{12} & \phi_{22} & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1p} & \phi_{2p} & \dots & \phi_{pp} \end{pmatrix} \begin{pmatrix} \phi_{11} & \phi_{12} & \dots & \phi_{1p} \\ 0 & \phi_{22} & \dots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \phi_{pp} \end{pmatrix}
\end{aligned}$$

Now suppose $j < k$

$$\begin{aligned}
\Omega_{kl} &= \Phi_{k.}(\Phi^T)_{.l} \\
&= (\phi_{k1}, ..., \phi_{kk}, 0, ..., 0) \begin{pmatrix} (\phi^T)_{1j} \\ \vdots \\ (\phi^T)_{jj} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \sum_{i=1}^{min(k,j)} \phi_{ki}(\phi^T)_{ij} \\
&= \sum_{i=1}^{min(k,j)} \phi_{ki}\phi_{ji} \\
&= \sum_{i=1}^{j} \phi_{ki}\phi_{ji} \\
&= \phi_{kj}\phi_{jj} + \sum_{i=1}^{j-1} \phi_{ki}\phi_{ji} = 0 \quad \text{as } (j,k) \notin E \\
\iff \phi_{kj}\phi_{jj} &= -(\sum_{i=1}^{j-1} \phi_{ki}\phi_{ji}) \\
\iff \phi_{kj} &= -\frac{1}{\phi_{jj}}(\sum_{i=1}^{j-1} \phi_{ki}\phi_{ji})
\end{aligned}$$

Thus if $j < k, (j,k) \notin E$, then ϕ_{kj} is in fact a polynomial in terms of $\{\phi_{j,1}, ..., \phi_{j,j-1}, \frac{1}{\phi_{jj}}, \phi_{k1}, ..., \phi_{k,j-1}\} \subset \{\phi_{ij}\}_{i \leq j, (i,j) \in E}$. To make this rigourous we simply proceed by induction by showing that the above argument holds for the first $j < k, (j,k) \notin E$. Then suppose it holds for the n-th such edge and show for the $(n+1)$-th. □

Note that part 3 of Lemma 12 implies that

$$\begin{aligned}\pi_{U,\delta}(\Phi) &= \pi_{U,\delta}(\Omega(\Phi))|\frac{\partial\Omega}{\partial\Phi}| \\ &\propto e^{-\frac{1}{2}tr(\Phi^T\Phi U)+\frac{\delta}{2}\log|\Phi^T\Phi|}\prod_{i=1}^{p}\phi_{ii}^{v_i+1} \\ &= e^{-\frac{1}{2}tr(\Phi^T\Phi U)+\frac{\delta}{2}\log|\Phi^T\Phi|\sum_{i=1}^{p}v_i+1\log\phi_{ii}} \\ &= e^{-\frac{1}{2}tr(\Phi^T\Phi U)+(\delta+v_i+1)\sum_{i=1}^{p}\log\phi_{ii}}\end{aligned}$$

The last step follows from the fact that Φ is upper triangular, which in turn implies that

$$\log|\Phi^T\Phi| = \log|\Phi||\Phi| = 2\log|\Phi| = 2\log\prod_{i=1}^{p}\phi_{ii} = 2\sum_{i=1}^{p}\log\phi_{ii}$$

Also note that as U is positive definite, it also has a Cholesky decomposition

$$\begin{aligned}U^{-1} &:= T^TT \\ \text{and } \Psi &:= \Phi T^{-1} \\ \implies tr(\Phi^T\Phi U) &= tr(\Phi^T\Phi T^{-1}(T^T)^{-1}) \\ &= tr(\Phi T^{-1}(T^T)^{-1}\Phi^T) \\ &= tr(\Phi T^{-1}(\Phi T^{-1})^T) \\ &= tr(\Psi\Psi^T)\end{aligned}$$

Note that as T is upper triangular, T^{-1} is also upper triangular. In addition, Φ is upper triangular. Hence, Ψ is upper triangular.

Lemma 13. *The mapping*

$$\{\Phi\}_{i\leq j,(i,j)\in E}\mapsto\{\Psi\}_{i\leq j,(i,j)\in E}$$

is a bijection from $\mathbb{R}^{|E|}\times\mathbb{R}_+^p$ to $\mathbb{R}^{|E|}\times\mathbb{R}_+^p$. This is clear as $\Psi = \Phi T^{-1}$ and T^{-1} is invertible. Secondly the Jacobian, informally can be though of as a vectorized version of $|\frac{\partial\Phi(\Psi)}{\partial\Psi}| = |\frac{\partial}{\partial\Psi}(T\Psi)| =$ function of T, which does not depend on Ψ.

Now note that $tr(AB) = \sum_{i=1}^{p}(AB)_{ii} = \sum_{i=1}^{p}\sum_{k=1}^{p}(A)_{ik}(B)_{ki} \implies tr(AA^T) = \sum_{i=1}^{p}\sum_{k=1}^{p}(A)_{ik}(A^T)_{ki} = \sum_{i=1}^{p}\sum_{k=1}^{p}(A)_{ik}^2$. As $i > j \implies \psi_{ij} =$

0, we have that $tr(\Psi\Psi^T) = \sum_{i=1}^p \sum_{k=1}^p \psi_{ik}^2 = \sum_{i=1}^p \sum_{k\geq i} \psi_{ik}^2$ Thus we can say that

$$\begin{aligned}
\pi_{U,\delta}(\Phi) &\propto e^{-\frac{1}{2}tr(\Phi^T\Phi U)+(\delta+v_i+1)\sum_{i=1}^p \log\phi_{ii}} \\
&= e^{-\frac{1}{2}tr(\Psi\Psi^T)+(\delta+v_i+1)\sum_{i=1}^p \log\psi_{ii}} \qquad \underbrace{|\frac{\partial\Phi(\Psi)}{\partial\Psi}|}_{\text{independent of }\Psi} \\
&\propto e^{-\frac{1}{2}\sum_{i=1}^p \sum_{i\leq k}\psi_{ik}^2+(\delta+v_i+1)\sum_{i=1}^p \log\psi_{ii}} \\
&\propto (\prod_{i=1}^p e^{-\frac{1}{2}\sum_{i\leq k}\psi_{ik}^2})(\prod_{i=1}^p (\psi_{ii}^2)^{\frac{\delta+v_i+1}{2}}) \\
&= (\prod_{i=1}^p e^{-\frac{1}{2}\sum_{i<k,(k,i)\notin E}\psi_{ik}^2-\frac{1}{2}\sum_{i<k,(k,i)\in E}\psi_{ik}^2-\frac{1}{2}\sum_{i=k}\psi_{ik}^2})(\prod_{i=1}^p (\psi_{ii}^2)^{\frac{\delta+v_i+1}{2}}) \\
&= \prod_{i=1}^p \Big(\underbrace{e^{-\frac{1}{2}\sum_{k=i+1,(k,i)\notin E}^p \psi_{ik}^2}}_{\text{uniformly bounded by 1}} \times \underbrace{e^{-\frac{1}{2}\sum_{k=i+1,(k,i)\in E}^p \psi_{ik}^2}}_{\psi_{ik}\sim\mathcal{N}(0,1)} \times \underbrace{(\psi_{ii}^2)^{\frac{\delta+v_i+1}{2}} e^{-\frac{1}{2}\psi_{ii}^2}}_{\psi_{ii}^2\sim\Gamma(\frac{\delta+v_i+1}{2},\frac{1}{2})}\Big) \\
&\leq \prod_{i=1}^p \Big((\prod_{k=i+1}^p \mathcal{N}(0,1))\times\Gamma(\frac{\delta+v_i+1}{2},\frac{1}{2})\Big)
\end{aligned}$$

This is in perfect form to apply the following Accept-Reject Algorithm with $M = 1$, $g = \prod_{i=1}^p \Big((\prod_{k=i+1}^p \mathcal{N}(0,1))\times\Gamma(\frac{\delta+v_i+1}{2},\frac{1}{2})\Big)$ when $f = \pi_{U,\delta}(\Phi)$.

Theorem 23. ***(Accept-Reject Algorithm)*** *Suppose the following hold:*

1. *Let X be a sample from the density g with support $\mathcal{X}$ and U is sample from the standard uniform distribution independent of g, i.e. $X \sim g$ and $U \sim \mathcal{U}(0,1)$*
2. *Let $R := \frac{f(X)}{Mg(X)}$, where $M \geq \sup_{x\in\mathcal{X}} \frac{f(x)}{g(x)}$*
3. *If $U < R$, let $Y = X$, i.e. accept Y as a sample from f. Else if $u \geq r$, reject x and return to Step 1.*

Then $Y \sim f$.

Proof. We need to show that $Y|(U < R) \sim f$. This is equivalent to showing that $P(Y \leq y|U < R) = F(y)$ where F is the cdf of f. First note that

$$\begin{aligned}
P(U \leq R) &= P(U \leq \frac{f(X)}{Mg(X)}) \\
&= \int_{-\infty}^{\infty} P(U \leq \frac{f(x)}{Mg(x)}|X = x)g(x)dx \\
&= \int_{-\infty}^{\infty} P(U \leq \frac{f(x)}{Mg(x)})g(x)dx \quad \text{as } U \perp\!\!\!\perp X \\
&= \int_{-\infty}^{\infty} \frac{f(x)}{Mg(x)}g(x)dx \quad \text{as } \forall u \in [0,1], P(U \leq u) = u \\
&= \int_{-\infty}^{\infty} \frac{f(x)}{M}dx \\
&= \frac{1}{M}.
\end{aligned}$$

Thus,

$$\begin{aligned}
P(Y \leq y | U \leq R) &= \frac{P(X \leq y, U \leq R)}{P(U \leq R)} \qquad \text{as } U \leq R \implies X = Y \\
&= \frac{P(U \leq R | X \leq y)P(X \leq y)}{P(U \leq R)} \\
&= \frac{G(y)}{P(U \leq R)} P(U \leq R | X \leq y) \\
&= \frac{G(y)}{P(U \leq R)} \frac{P(U \leq \frac{f(X)}{Mg(X)}, X \leq y)}{P(X \leq y)} \\
&= \frac{G(y)}{P(U \leq R)} \frac{\int_{-\infty}^{y} P(U \leq \frac{f(w)}{Mg(w)}, X = w)g(w)dw}{G(y)} \\
&= \frac{G(y)}{P(U \leq R)} \frac{\int_{-\infty}^{y} P(U \leq \frac{f(w)}{Mg(w)})g(w)dw}{G(y)} \qquad \text{as } U \perp\!\!\!\perp X \\
&= \frac{1}{P(U \leq R)} \int_{-\infty}^{y} \frac{f(w)}{Mg(w)} g(w)dw \qquad \text{as } \forall u \in [0,1], P(U \leq u) = u \\
&= \frac{1}{P(U \leq R)} \int_{-\infty}^{y} \frac{f(w)}{Mg(w)} g(w)dw \\
&= \frac{1}{P(U \leq R)} \frac{F(y)}{M} \\
&= F(y).
\end{aligned}$$

□

6.13 Independence Metropolis Algorithm

This is due to Mitsakakis, et. al (2011) in EJS. Recall that in any independence metropolis algorithm, we want to generate a sample $\pi(.)$ on a sample space $\mathcal{X}$ and we have another density $q(.)$ with support $\mathcal{X}$, which is easy to sample from.

Definition 24. *A probability distribution T has stationary distribution Π if*

$$\forall C \in \mathcal{B}, \quad \Pi(C) = \int_C T(y, C)\Pi(dy)$$

Note that we have used the same notation for the probability measure induced by a distribution function and the distribution function in the above definition.

Definition 25. *A probability distribution, T, with probability density, t, is reversible with respect to Π with density π if and only if*

$$\pi(x)t(x,y) = \pi(y)t(y,x)$$

Intuitively this says that the probability of being at x and then moving to y is equal to the probaility of being at y and then moving to x.

Lemma 14. *If a probability measure, T, with probability density, t, is reversible with respect to Π with density π, then the stationary distribution of T is Π.*

Theorem 26. *Suppose $\Pi(.)$ has density $\pi(.)$ and $Q(.)$ has density $q(.)$. Also suppose we generate X_r according to the following steps:*

1. *Sample initial $X_0 \sim q(.)$ and set $r = 0$.*
2. *Generate $\tilde{X}_{r+1} \sim q(.)$*
3. *Compute $\alpha(X_r, \tilde{X}_{r+1}) = \min\{1, \frac{\pi(\tilde{X}_{r+1})/q(\tilde{X}_{r+1})}{\pi(X_r)/q(X_r)}\}$*
4. *Generate $U \sim \mathcal{U}(0,1)$*
5. *If $U < \alpha(X_r, \tilde{X}_{r+1})$, set $X_{r+1} = \tilde{X}_{r+1}$. Else if $U \geq \alpha$, set $X_{r+1} = X_r$.*
6. *Set $r = r + 1$ and go to Step 2.*

Then as $r \to \infty, X_r \sim \Pi(.)$ in the sense that

$$\lim_{r\to\infty} \sup_{C\in\mathcal{B}} |P(X_r \leq C) - \pi(C)| \to 0$$

and

$$P\big(\lim_{r\to\infty} \sum_{i=1}^{r} h(X_r) = \int h(y)\pi(y)dy\big) = 1.$$

We can use the Lemma 14 to show that X_r generated according to Theorem 26 has Π as it's stationary distribution. What this means is that if $X_r \sim \pi$, then $X_{r+1} \sim \pi$. At the outset note, given $\tilde{X}_{r+1}$ that $P(X_{r+1} = X_r | X_r = x) = P(U > \alpha(x, \tilde{X}_{r+1})) = 1 - \alpha(x, \tilde{X}_{r+1})$ and $P(X_{r+1} = \tilde{X}_{r+1} | X_r = x) = \alpha(x, \tilde{X}_{r+1})$. Hence, given $\tilde{X}_{r+1}$, we can see intuitively that

$$\begin{aligned}
P(X_{r+1} \in C | X_r = x) &= P(X_r \in C | (X_{r+1} = X_r | X_r = x)) \\
&\quad \times P(X_{r+1} = X_r | X_r = x) \\
&+ P(\tilde{X}_{r+1} \in C | (X_{r+1} = \tilde{X}_{r+1} | X_r = x)) \\
&\quad \times P(X_{r+1} = \tilde{X}_{r+1} | X_r = x) \\
\implies P(X_{r+1} \leq y | X_r = x) &= P(X_r \leq y | X_r = x) \\
&\quad \times P(X_{r+1} = X_r | X_r = x) \\
&+ P(\tilde{X}_{r+1} \leq y | (X_{r+1} = \tilde{X}_{r+1} | X_r = x)) \\
&\quad \times P(X_{r+1} = \tilde{X}_{r+1} | X_r = x) \\
\implies \frac{d}{dy} P(X_{r+1} \leq y | X_r = x) &= \frac{d}{dy} (P(X_r \leq y | X_r = x) \delta_{X_r}(X_{r+1}) \\
&\quad \times P(X_{r+1} = X_r | X_r = x)) \\
&+ \frac{d}{dy} (P(\tilde{X}_{r+1} \leq y | X_r = x) \delta_{\tilde{X}_{r+1}}(X_{r+1}) \\
&\quad \times P(X_{r+1} = \tilde{X}_{r+1} | X_r = x)) \\
\implies f_{X_{r+1}}(y | X_r = x) &= f_{X_r}(y | X_r = x)(1 - \alpha(x, \tilde{X}_{r+1})) \delta_{X_r}(X_{r+1}) \\
&+ f_{\tilde{X}_{r+1}}(y | X_r = x) \alpha(x, \tilde{X}_{r+1}) \delta_{\tilde{X}_{r+1}}(X_{r+1}) \\
\implies f_{X_{r+1}}(y | X_r = x) &= f_{X_r}(x)(1 - \alpha(x, \tilde{X}_{r+1})) \delta_x(y) \\
&+ q(x, y) \alpha(x, y)
\end{aligned}$$

this part is a little strange. need to fix it up

Proof. More formally, we can write the probability of transitioning from x to dy as $P(x, dy) = \gamma(x)\delta_x(dy) + p(x, dy) = \gamma(x)\delta_x(dy) + p(x, y)\mu(dy)$ where

$$p(x, y) := \begin{cases} \alpha(x, y) q(x, y) & \text{if } y \neq x \\ 0 & \text{if } y = x \end{cases}$$

and

$$\gamma(x) = 1 - \int_{\mathcal{Y}} \alpha(x,y)q(x,dy)$$

and if we can show that $\pi(x)\gamma(x)\delta_x(dy) = \pi(y)\gamma(y)\delta_y(dx)$ and $\pi(x)p(x,y)\mu(dy) = \pi(y)p(y,x)\mu(dy)$, we are done. The above formulation is more general than the statement of the Theorem which only talks about the case where $q(x,y) = q(y)$. First lets focus on the second term above. We want to show that $\pi(x)p(x,y) = \pi(y)p(y,x)$. If $x = y$, then $p(x,y) = 0 = p(y,x)$. Thus $\pi(x)p(x,y) = \pi(y)p(y,x)$ If $x \neq y$ we need to consider two cases:

1. $\pi(x)q(x,y) > \pi(y)q(y,x) \implies \alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$ and $\alpha(y,x) = 1$. Then

$$\begin{aligned} \pi(x)p(x,y) &= \pi(x)q(x,y)\alpha(x,y) \\ &= \pi(x)q(x,y)\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \\ &= \pi(y)q(y,x) \\ &= \pi(y)q(y,x)\alpha(y,x) \\ &= \pi(y)p(y,x). \end{aligned}$$

2. $\pi(x)q(x,y) > \pi(y)q(y,x)$ is a similar case.

Now consider the first part: $\pi(x)\gamma(x)\delta_x(dy)$. Note that both sides are equal to 0 if $x \neq y$ and in the case $x = y$ we have that $\pi(x)\gamma(x)\delta_x(dy) = \pi(y)\gamma(y)\delta_y(dx)$. □

We want an independence metropolis algorithm for the G-Wishart where

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

What makes it difficult to sample is the $\Omega \in \mathbb{P}_G$ constraint. Recall the transformation from the previous section

$$\begin{aligned} \Omega &= \Phi^T\Phi \\ U^{-1} &= T^T T \\ \Psi &= \Phi T^{-1} \end{aligned}$$

where Φ, T and Ψ are all upper triangular with postive diagonal entries. Now the transformation $\{\Omega_{ij}\}_{i \leq j,(i,j) \in E} \mapsto \{\psi_{ij}\}_{i \leq j,(i,j) \in E}$ gives us the following density

$$\pi_{u,\delta}(\Psi) = C_1(\prod_{i=1}^{p} e^{-\frac{1}{2}\sum_{i<k,(k,i)\notin E}\psi_{ik}^2 - \frac{1}{2}\sum_{i<k,(k,i)\in E}\psi_{ik}^2})(\prod_{i=1}^{p}(\psi_{ii}^2)^{\frac{\delta+v_i+1}{2}})e^{-\frac{1}{2}\sum_{i=1}^{p}\psi_{ii}^2}$$

Let

$$\begin{aligned} q_{U,\delta}(\Psi) &= C_2(\prod_{i=1}^{p} e^{-\frac{1}{2}\sum_{i<k,(k,i)\in E}\psi_{ik}^2})(\prod_{i=1}^{p}(\psi_{ii}^2)^{\frac{\delta+v_i+1}{2}})e^{-\frac{1}{2}\sum_{i=1}^{p}\psi_{ii}^2} \\ &= \prod_{i=1}^{p}\Big((\prod_{k=i+1}^{p}\mathcal{N}(0,1)) \times \Gamma(\frac{\delta+v_i+1}{2},\frac{1}{2})\Big) \end{aligned}$$

and we can easily generate from this density. Then

$$\begin{aligned} \frac{\pi_{u,\delta}(\Psi)}{q_{U,\delta}(\Psi)} &= \frac{C_1}{C_2}(\prod_{i=1}^{p} e^{-\frac{1}{2}\sum_{i<k,(k,i)\notin E}\psi_{ik}^2}) \\ \implies \quad \frac{\pi_{u,\delta}(\Psi^{r+1})/q_{U,\delta}(\Psi^{r+1})}{\pi_{u,\delta}(\Psi^{r})/q_{U,\delta}(\Psi^{r})} &= \frac{(\prod_{i=1}^{p} e^{-\frac{1}{2}\sum_{i<k,(k,i)\notin E}(\psi_{ik}^{r+1})^2})}{(\prod_{i=1}^{p} e^{-\frac{1}{2}\sum_{i<k,(k,i)\notin E}(\psi_{ik}^{r})^2})}, \end{aligned}$$

that is, the normalizing constants in the ratio cancel out. Hence we can use the following algorithm

1. Generate $\Psi^0 \sim q_{U,\delta}(.)$ and let $r = 0$.
2. Generate $\tilde{\Psi}^{r+1} \sim q_{U,\delta}(.)$.
3. Compute

$$\begin{aligned} \alpha &= \min\{1, \frac{\pi(\tilde{\Psi}^{r+1})/q(\tilde{\Psi}^{r+1})}{\pi(\Psi^{r+1})/q(\Psi^{r+1})}\} \\ &= \min\{1, \prod_{i=1}^{p} e^{-\frac{1}{2}\sum_{i<k,(k,i)\notin E}(\tilde{\psi}_{ik}^{r+1})^2 + \frac{1}{2}\sum_{i<k,(k,i)\notin E}(\psi_{ik}^{r})^2}\} \end{aligned}$$

   For numerical stability it is advisable to calculate the exponent before taking the exponential.

4. Generate $U \sim \mathcal{U}(0,1)$. Set
$$\Psi^{r+1} = \begin{cases} \tilde{\Psi}^{r+1} & \text{if } U < \alpha \\ \Psi^{r} & \text{if } U \geq \alpha \end{cases}$$

5. Set $r = r+1$ and repeat until convergence.

It is important to consider what we mean by convergence. Recall that $tr(\Omega U) = \sum_{i=1}^{p} (\Omega U)_{ii} = \sum_{i=1}^{p} \Omega_{i.} U_{.i} = \sum_{i=1}^{p} \sum_{j=1}^{p} \Omega_{ij} U_{ji} \implies \frac{d}{d\Omega_{i,j}} tr(\Omega U) = U_{ji} = U_{ij}$. Also recall that $\frac{d}{d\Omega}|\Omega| = |\Omega|(\Omega^{-1})^T = |\Omega|\Sigma^T = |\Omega|\Sigma \implies \frac{d}{d\Omega} log|\Omega| = \frac{1}{|\Omega|}\frac{d}{d\Omega}|\Omega| = \Sigma$. Note that the normalizing constant for the G-Wishart, $Z_G(U,\delta)$

$$\begin{aligned} Z_G(U,\delta) &= \int_{\mathbb{P}_G} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} d\Omega, \\ \iff 1 &= \frac{1}{Z_G(U,\delta)} \int_{\mathbb{P}_G} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} d\Omega, \\ \implies \frac{d}{\Omega_{i,j}} 1 &= \frac{d}{d\Omega_{i,j}} \frac{1}{Z_G(U,\delta)} \int_{\mathbb{P}_G} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} d\Omega \end{aligned}$$

Thus for $(i,j) \in E$ or $i = j$

$$\begin{aligned} 0 &= \int_{\mathbb{P}_G} \frac{d}{d\Omega_{i,j}} \frac{1}{Z_G(U,\delta)} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} d\Omega \\ \iff 0 &= \int_{\mathbb{P}_G} \frac{1}{Z_G(U,\delta)} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} \frac{d}{d\Omega_{i,j}} (-\frac{1}{2} tr(\Omega U) + \frac{\delta}{2}\log|\Omega|) d\Omega \\ \iff 0 &= \int_{\mathbb{P}_G} \frac{1}{Z_G(U,\delta)} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} \frac{d}{d\Omega_{i,j}} (-\frac{1}{2} tr(\Omega U) + \frac{\delta}{2}\log|\Omega|)(-\frac{1}{2}U_{i,j} + \frac{\delta}{2}\Sigma_{i,j}) d\Omega \end{aligned}$$

which implies that,

$$\begin{aligned} &\int_{\mathbb{P}_G} \frac{U_{i,j}}{\delta} \frac{1}{Z_G(U,\delta)} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} \frac{d}{d\Omega_{i,j}} (-\frac{1}{2} tr(\Omega U) + \frac{\delta}{2}\log|\Omega|) d\Omega \\ &= \int_{\mathbb{P}_G} \Sigma_{i,j} \frac{1}{Z_G(U,\delta)} e^{-\frac{1}{2}tr(\Omega U) + \frac{\delta}{2}\log|\Omega|} \frac{d}{d\Omega_{i,j}} (-\frac{1}{2} tr(\Omega U) + \frac{\delta}{2}\log|\Omega|) d\Omega \end{aligned}$$

Thus $\frac{U_{i,j}}{\delta} = \mathbb{E}[\Sigma_{i,j}]$ for $(i,j) \in E$ or $i = j$. From Theorem 26 we have that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \Sigma_{i,j}^{k} = E[\Sigma_{i,j}] = \frac{U_{i,j}}{\delta}, \quad a.s.$$

Therefore one criterion for convergence could be if for any pre-specified $\epsilon > 0$, we decide that convergence has happened if $\sup_{(i,j)\in E}|\frac{1}{k}\sum_{i=1}^{k}\Sigma_{i,j}^{k} - \frac{U_{i,j}}{\delta}| < \epsilon$.

7 Covariance Graph Models

We want to estimate Σ where $Y^1, ..., Y^n \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$. We are especially interested in the case when $n < p$. Recall that Σ has be be positive definite. As $Y_iY_i^T$ is a $p \times p$ matrix of rank 1, $S = \frac{1}{n}\sum_{i=1}^{n} Y_iY_i^T$ can be full rank and hence positive definite if $n \geq p$. If $Y_1 \perp\!\!\!\perp Y_2 \implies Y_1$ is linearly independent of Y_2 with probability 1. To see this intuitively, note that Y_1 linearly dependent on Y_2 implies that $Y_1 = cY_2 \implies P(Y_1 - cY_2 = 0) = 1$. But this is impossible as $Y_1, Y_2 \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma) \implies Y_1 - cY_2 \sim \mathcal{N}_p(0, (1 + c^2)\Sigma)$. Thus, with probability 1, we can say that Y^1 and Y^2 are linearly independent. As a result 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 $S = \frac{1}{n}\sum_{i=1}^{n} Y_iY_i^T$ to find our positive definite estimate, $\hat{\Sigma}$, of Σ.

7.1 Iterated Conditional Fitting (ICF)

Recall that in Section 6.4 we showed that we can derive a closed form expression for Ω_{mle} when $n > \max_{i=1,..,k}|C_i|$ where k is the number of cliques. However, Σ_{mle} doesn't even exist when $n < p$. Now suppose that

$$l^*(\Sigma) = tr(\Sigma^{-1}S) + log|\Sigma|, \quad \Sigma \in \mathbb{P}_G.$$

It is clear that when $n \geq p$ and S is invertible, we can minimize $l^*(\Sigma)$ as defined above by setting $\hat{\Sigma}_{mle} = S$ However, note that for $n < p$, S is not invertible and hence $S \notin \mathbb{P}_G$ as a result cannot be used as an *mle* estimate for Σ. Another problem we run into is that $l^*(\Sigma)$ is not convex as a function of Σ, unlike $l^*(\Omega)$ which was convex as a function of Ω. Chaudhuri, Drton and Richardson outline the following algorithm, called Iterated Conditional Fitting (ICF), which is a partial minimization procedure. However, it's convergence is not guaranteed. The basic idea is outlined below. Let

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix} = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{-1,1}^T \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix},$$

$$S = \begin{pmatrix} S_{1,1} & S_{-1,1}^T \\ S_{-1,1} & S_{-1,-1} \end{pmatrix}$$

and

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\gamma_1} & -\frac{1}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1} \\ -\frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1} & \Sigma_{-1,-1}^{-1} + \frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1} \end{pmatrix}$$

where $\gamma_1 = \Sigma_{1,1} - \Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}$ Then $\Sigma \mapsto (\Sigma_{-1,1}, \Sigma_{-1,-1}, \gamma_1)$ is a bijection. Thus as long as $l^*(\Sigma)$ has a unique global minimum Theorem 21 implies that

$$\arg\max_{\Sigma} l^*(\Sigma) = \arg\max_{\Sigma_{-1,1},\Sigma_{-1,-1},\gamma_1} l^*(\Sigma_{-1,1}, \Sigma_{-1,-1}, \gamma_1).$$

Thus once we have found $(\Sigma_{-1,1}, \Sigma_{-1,-1}, \gamma_1)$ from Σ we need to minimize $l^*(\Sigma)$. Note that

$$\begin{aligned}
l^*(\Sigma_{-1,1}) &= tr(\Sigma^{-1}S) + \log|\Sigma| \\
&= tr((\Sigma^{-1}S)_{1,1} + (\Sigma^{-1}S)_{-1,-1}) + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= (\Sigma^{-1}S)_{1,1} + tr((\Sigma^{-1}S)_{-1,-1}) + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,1} \\
&\quad + tr(-\frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}S^T_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1}S_{-1,-1}\frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,1} \\
&\quad + tr(-\frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}S^T_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1}S_{-1,-1} + \frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,1} \\
&\quad + tr(-\frac{1}{\gamma_1}S^T_{-1,1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1}S_{-1,-1} + \frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,1} \\
&\quad - \frac{1}{\gamma_1}S^T_{-1,1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1} \\
&\quad + tr(\Sigma^{-1}_{-1,-1}S_{-1,-1} + \frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1
\end{aligned}$$

Thus

$$\begin{aligned}
l^*(\Sigma_{-1,1}) &= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1}S_{-1,-1} + \frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1}S_{-1,-1}) + tr(\frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1}S_{-1,-1}) + tr(\frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1}S_{-1,-1}) + \frac{1}{\gamma_1}tr(\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}) \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1}S_{-1,-1}) + \frac{1}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1} \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1.
\end{aligned}$$

In the last line we have used the fact that

$$tr(\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}) = \Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}$$

for $\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}S_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1} \in \mathbb{R}$. Note that this is also a quadratic form in $\Sigma_{-1,1}$.

Lemma 15. *Let*

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{-1,1}^T \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix}$$

and $\gamma_1 = \Sigma_{1,1} - \Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} \Sigma_{-1,1}$. *Then regardless of any choice of* $\Sigma_{-1,1}$, Σ *is positive definite if* $\Sigma_{-1,-1}$ *is positive definite and* $\gamma_1 > 0$.

As a result, we get positive definiteness of Σ so long as $\Sigma_{-1,-1}$ is positive definite and $\gamma_1 > 0$ and only need to work abpout the 0 constraints in $\Sigma_{-1,1}$. In this respect define B_1 to be the non-zero entires of $\Sigma_{-1,1}$.

$$B_1 \coloneqq ((\Sigma_{-1,1})_{i,j})_{(i,j)\in E}$$

and Q_1 be the $|\Sigma_{-1,1}| \times |B_1|$ matrix such that $Q_1 B_1 = \Sigma_{-1,1}$, which essentially requires that the i-th column of Q_1 will have a 1 in the j-th row if the j-th entry of $\Sigma_{-1,1} \neq 0$.

Example As an example consider

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & \sigma_{13} & 0 & \sigma_{15} \\ 0 & \sigma_{22} & 0 & 0 & \sigma_{25} \\ \sigma_{31} & 0 & \sigma_{33} & 0 & \sigma_{35} \\ 0 & 0 & 0 & \sigma_{45} & 0 \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & 0 & \sigma_{55} \end{pmatrix}.$$

Then

$$\Sigma_{-1,1} = \begin{pmatrix} 0 \\ \sigma_{31} \\ 0 \\ \sigma_{51} \end{pmatrix},$$

which implies that

$$B_1 = \begin{pmatrix} \sigma_{31} \\ \sigma_{51} \end{pmatrix} \text{ and } Q_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Then

$$\begin{aligned}
l^*(B_1) &= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}B_1^T Q_1^T \Sigma_{-1,-1}^{-1} S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1} S_{-1,-1}) + \frac{1}{\gamma_1} B_1^T Q_1^T \Sigma_{-1,-1}^{-1} S_{-1,-1} \Sigma_{-1,-1}^{-1} Q_1 B_1 \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}(B_1^T Q_1^T \Sigma_{-1,-1}^{-1} S_{-1,-1} \Sigma_{-1,-1}^{-1} Q_1 B_1 \\
&\quad - 2B_1^T Q_1^T \Sigma_{-1,-1}^{-1} S_{-1,1}) \\
&\quad + \varsigma(\gamma_1, \Sigma_{-1,-1})
\end{aligned}$$

Now to find the minimum point note that if $f(x) = x^T B x - 2x^T A + c$, then $\frac{d}{dx} f(x) = (B + B^T)x - 2A^T \overset{set}{=} 0 \implies x = (B + B^T)^{-1} 2A^T$. If $B = B^T$, then $x = (2B)^{-1} 2A^T = B^{-1} A^T$. Thus

$$\hat{B}_1 = (Q_1^T \Sigma_{-1,-1}^{-1} S_{-1,-1} \Sigma_{-1,-1}^{-1} Q_1)^{-1} (Q_1^T \Sigma_{-1,-1}^{-1} S_{-1,1})^T$$

which looks somewhat like a solution to a regression problem. This is essentailly how we minimize with respect to $\Sigma_{-1,1}$.

Now to minimize with respect to γ_1 holding $\Sigma_{-1,1}$ and $\Sigma_{-1,-1}$ fixed, note that

$$\begin{aligned}
l^*(\gamma_1) &= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} S_{-1,1} \\
&\quad + tr(\Sigma_{-1,-1}^{-1} S_{-1,-1}) + \frac{1}{\gamma_1}\Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} S_{-1,-1} \Sigma_{-1,-1}^{-1} \Sigma_{-1,1} \\
&\quad + \log|\Sigma_{-1,-1}|\gamma_1 \\
&= \frac{1}{\gamma_1}S_{1,1} - \frac{2}{\gamma_1}\Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} S_{-1,1} \\
&\quad + \frac{1}{\gamma_1}\Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} S_{-1,-1} \Sigma_{-1,-1}^{-1} \Sigma_{-1,1} \\
&\quad + \log\gamma_1 + \varsigma(\Sigma_{-1,1}, \Sigma_{-1,-1}) \\
&= \frac{1}{\gamma_1}(S_{1,1} - 2\Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} S_{-1,1} \\
&\quad + \Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} S_{-1,-1} \Sigma_{-1,-1}^{-1} \Sigma_{-1,1}) \\
&\quad + \log\gamma_1 + \varsigma(\Sigma_{-1,1}, \Sigma_{-1,-1})
\end{aligned}$$

Now note that to minimize $f(x) = \frac{a}{x} + \log(x) + c$ we can take the derivative $\frac{d}{dx}f(x) = -\frac{a}{x^2} + \frac{1}{x} \overset{set}{=} 0 \implies \frac{a}{x^2} = \frac{1}{x} \iff x = a$. Thus

$$\hat{\gamma_1} = S_{1,1} - 2\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,1} + \Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}S_{-1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}$$

In the above formulation Σ and S maintains positive definite-ness we have that $\hat{\gamma_1} > 0$, which is what we need for Σ to be positive definite. Finally note that it is not possible to minimize $l^*(\Sigma_{-1,-1})$. Hence the ICF algorithm consists of the following steps:

1. Set $i = 1$. Decompose $\Sigma \mapsto (\Sigma_{-i,i}, \Sigma_{-i,-i}, \gamma_i)$, which is a bijection.
2. Minimize $l^*(\Sigma)$ with respect to $\Sigma_{-i,i}$ holding $(\Sigma_{-i,-i}, \gamma_i)$ constant
3. Minimize $l^*(\Sigma)$ with respect to γ_i holding $(\Sigma_{-i,-i}, \Sigma_{-i,i})$ constant
4. Repeat for $i = 2, ..., p$.

However, there is no guarantee that this algorithm will indeed converge. In practice, it usually converges but due to non-convexity of $l^*(\Sigma)$ in terms of Σ, it may not converge to a global minimum. As a result, this should implemented from multiple starting values and we should pick the best Σ which gives us the lowest value of $l^*(\Sigma)$.

Now define T_i to be the minimization in ICF with respect to $\Sigma_{-i,i}$ holding $(\Sigma_{-i,-i}, \gamma_i)$ constant,

$$T_i(\Sigma) = \arg \min_{\tilde{\Sigma} \in \mathbb{P}_G, \tilde{\Sigma}_{-i,-i} = \Sigma_{-i,-i}, \gamma_i(\tilde{\Sigma}) = \gamma_i(\Sigma)} l^*(\tilde{\Sigma})$$

and R_i to be the minimization in ICF γ_i holding $(\Sigma_{-i,-i}, \Sigma_{-i,i})$ constant

$$R_i(\Sigma) = \arg \min_{\tilde{\Sigma} \in \mathbb{P}_G, \tilde{\Sigma}_{-i,-i} = \Sigma_{-i,-i}, \tilde{\Sigma}_{-i,i} = \Sigma_{-i,i}} l^*(\tilde{\Sigma})$$

where

$$l^*(\Sigma) = tr(\Sigma^{-1}S) + \log|\Sigma|, \quad \Sigma \in \mathbb{P}_G.$$

Note that we can redefine the ICF algorithm as

1. Select an initial value Σ^0 and set $r = 0$.
2. $\Sigma^{r+1} = R_pT_p...R_1T_1(\Sigma^r)$.

3. Repeat the previous step until convergence.

This algorithm produces a sequence of iterates $\{\Sigma^r\}_{r\geq 0}$.

Lemma 16. *If* $(x_n)_{n=1}^{\infty} \in K$ *where* K *is compact, then* $(x_n)_{n=1}^{\infty}$ *has a convergent subsequence, i.e.* $\exists (n_k) \subset (n)$ *such that* $x_{n_k} \to x \in K$*.*

Now define Θ_0 as the space consisting of Σ such that $l^*(\Sigma) \leq l^*(\Sigma^0)$

$$\Theta_0 = \{\Sigma \in \mathbb{P}^+ : l^*(\Sigma) \leq l^*(\Sigma^0)\}$$

Then the sequence produced by the ICF algorithm $\{\Sigma^r\}_{r\geq 0} \subset \Theta_0$.

Lemma 17. *(Heine-Borel Theorem) A subset of* $\mathbb{R}^m$ *is compact if and only if it is closed and bounded.*

Lemma 18. *If* $D \subset B$ *is a closed set where* $f : A \mapsto B$ *is continuous function on* A*, then* $f^{-1}(D)$ *is closed.*

Proof. Recall that the pre-image of an open set is open. Thus $f^{-1}(B - D)$ is open. Thus $f^{-1}(B-D) = f^{-1}(B \cap D^c) = f^{-1}(B) \cap f^{-1}(D^c) = A \cap f^{-1}(D^c) = A \cap (f^{-1}(D))^c$ is open, using the fact that the preimage of a complement is the complement of the pre-image. Thus $f^{-1}(D)$ is closed. □

Lemma 19. $\Theta_0 = \{\Sigma \in \mathbb{P}^+ : l^*(\Sigma) \leq l^*(\Sigma^0)\}$ *is a compact set.*

Proof. We will show that $\Theta_0 \subset \mathbb{P}^+ \subset \mathbb{R}^{p\times p}$ is closed and bounded. Note that $l^*(\Sigma)$ is a continuous function in terms of Σ and hence as $(-\infty, b]$ is a closed set, $(l^*)^{-1}((-\infty, b])$ is a closed set. In particular, $\Theta_0 = (l^*)^{-1}((-\infty, l^*(\Sigma^0)])$ is closed.

Now let

$$\Sigma = P\Lambda P^T$$

be the spectral decomposition of Σ and suppose that $\Lambda = diag(\lambda_1, ..., \lambda_p)$ where $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_p$. Then

$$|\Sigma| = |P\Lambda P^T| = |P||\Lambda||P| = |\Lambda|,$$

$$\Sigma^{-1} = P\Lambda^{-1}P^T,$$

where $\Lambda^{-1} = diag(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_p})$ and

$$\begin{aligned} l^*(\Sigma) &= tr(\Sigma^{-1}S) + \log|\Sigma| \\ &= tr(P\Lambda^{-1}P^T S) + \log|\Lambda| \\ &= tr(\Lambda^{-1}P^T SP) + \log|\Lambda| \\ &= tr(\Lambda^{-1})tr(P^T SP) + \log|\Lambda| \\ &= \sum_{i=1}^{p} \frac{1}{(\Lambda)_{ii}}(P^T SP)_{ii} + \sum_{i=1}^{p} \log(\Lambda)_{ii} \end{aligned}$$

and define

$$l_i^* = \frac{1}{\lambda_i}(P^T SP)_{ii} + \log \lambda_i$$

Note that l_i^* is finite unless $\lambda_i \to 0$ or $\lambda_i \to \infty$. If

$$\lambda_i \to 0 \implies \frac{1}{\lambda_i}(P^T SP)_{ii} \to \infty \implies l_i^* \to \infty$$

and

$$\lambda_i \to \infty \implies \log \lambda_i \to \infty \implies l_i^* \to \infty.$$

But this is impossible if $\Sigma \in \Theta_0$ and $|l^*(\Sigma^0)| < \infty$ as $l^*(\Sigma) \leq l^*(\Sigma^0), \forall \Sigma \in \Theta_0$. Finally note that $(P^T SP)_{ii} = P_{.i}^T SP_{.i} \geq 0$ as S is non-negative definite. Thus $l^*(\Sigma^0) \geq l^*(\Sigma) = \sum_{i=1}^p l_i^* \geq \sum_{i=1}^p \log \lambda_i$, which implies that $l^*(\Sigma)$ is bounded. And thus Θ_0 must be bounded as well,.i.e. $\exists \Sigma \in \mathbb{P}^+$ and $r > 0$ such that $d(\Theta_0, \Sigma) < r$. □

This guanrantees that Θ_0 has a global minimum, say Σ^*, not necessarily unique.

Lemma 20. *(Stationary points of ICF) If Σ is such that $T_i(\Sigma) = \Sigma$ and $R_i(\Sigma) = \Sigma$ for every $1 \leq i \leq p$, then Σ is a stationary point of l^*. i.e. $\frac{d}{d\Sigma} l^*(\Sigma) = 0$.*

Proof. Note that to minimize $l^*(\Sigma)$ with respect to $\Sigma_{-1,1}$ we need to differentiate $l^*(\Sigma)$ with respect to $\Sigma_{-1,1}$ and set the derivative equal to 0, i.e. $\frac{\partial l^*(\Sigma)}{\partial \Sigma_{-1,1}} = 0$. Similarly, to minimize $l^*(\Sigma)$ with respect to γ_1 we need to differentiate $l^*(\Sigma)$ with respect to γ_1 and set the derivative equal to 0, i.e. $\frac{\partial l^*(\Sigma)}{\partial \gamma_1} = 0$. Thus if for all $1 \leq i \leq p, T_i(\Sigma) = \Sigma$ and $R_i(\Sigma) = \Sigma$, then we must have that

$\frac{\partial l^*(\Sigma)}{\partial \Sigma_{-i,i}} = 0$ and $\frac{\partial l^*(\Sigma)}{\partial \gamma_i} = 0$. Note that if $\tilde{\Sigma} \in \{\tilde{\Sigma} \in \mathbb{P}_G, \tilde{\Sigma}_{-i,-i} = \Sigma_{-i,-i}, \tilde{\Sigma}_{-i,i} = \Sigma_{-i,i}\}$, then $\tilde{\Sigma}_{i,i} = \gamma_i(\tilde{\Sigma}) + \tilde{\Sigma}^T_{-i,i}\tilde{\Sigma}^{-1}_{-i,-i}\tilde{\Sigma}_{-i,i}$ is a linear function of $\gamma_1(\tilde{\Sigma})$ as $\tilde{\Sigma}^T_{-i,i}\tilde{\Sigma}^{-1}_{-i,-i}\tilde{\Sigma}_{-i,i}$ is fixed for $\tilde{\Sigma} \in \{\tilde{\Sigma} \in \mathbb{P}_G, \tilde{\Sigma}_{-i,-i} = \Sigma_{-i,-i}, \tilde{\Sigma}_{-i,i} = \Sigma_{-i,i}\}$, which implies that $\frac{\partial \gamma_i}{\partial \Sigma_{i,i}} = c$ where c is a constant. Thus $\frac{\partial l^*(\Sigma)}{\partial \Sigma_{i,i}} = \frac{\partial l^*(\Sigma)}{\gamma_i}\frac{\partial \gamma_i}{\partial \Sigma_{i,i}} = 0$. Thus $\frac{d}{d\Sigma}l^*(\Sigma) = 0$. □

Note that this lemma implies if we define as sections

$$\Theta_{i1}(\Sigma^*) = \{\tilde{\Sigma} \in \mathbb{P}_G, \tilde{\Sigma}_{-i,-i} = \Sigma_{-i,-i}, \tilde{\Sigma}_{-i,i} = \Sigma_{-i,i}\}$$
$$\Theta_{i2}(\Sigma^*) = \{\tilde{\Sigma} \in \mathbb{P}_G, \tilde{\Sigma}_{-i,-i} = \Sigma_{-i,-i}, \gamma_i(\tilde{\Sigma}) = \gamma_i(\Sigma)\}$$

for each $i = 1, ..., p$, then $T_i(\Sigma^*) = \arg\min_{\Sigma \in \Theta_{i1}(\Sigma^*)} l^*(\Sigma)$ and $R_i(\Sigma^*) = \arg\min_{\Sigma \in \Theta_{i2}(\Sigma^*)} l^*(\Sigma)$. Moreover, if we assume that Σ^* minimizes $l^*(\Sigma)$ over all sections, then $\frac{d}{d\Sigma}l^*\Sigma|_{\Sigma=\Sigma^*} = 0$. In fact, the lemma proves that this is true as well.

Lemma 21. *The sequence $\{l^*(\Sigma^r)\}_{r\geq 0}$ converges to a limit l_∞. Furthermore if $\Sigma^* \in A_\infty =$ set of accumulation points of $\{\Sigma^r\}_{r\geq 0}$, then $l^*(\Sigma^*) = l_\infty$ and $\frac{d}{d\Sigma}l^*(\Sigma)|_{\Sigma=\Sigma^*} = 0$.*

Proof. Note that by constrution $l^*(\Sigma^r)$ is monotonic and bounded. Thus it must converge to a limit, say l_∞. Now if $\Sigma^* \in A_\infty$, then $\exists \Sigma^{r_k}$ such that $\lim_{k\to\infty} \Sigma^{r_k} = \Sigma^*$, which implies that $\lim_{k\to\infty} l^*(\Sigma^{r_k}) = l^*(\Sigma^r)$. Next define $S := T_pR_p...T_1R_1$. Then $S(A_\infty)$ is the set of accumulation points of $S(\Sigma^r) = \Sigma^{r+1}$, which implies that $S(A_\infty) = A_\infty$ which implies that $l^*(S(A_\infty)) = l_\infty$. Then using the fact that T_i and R_i are all partial minimizations, for any $\Sigma^* \in A_\infty$

$$\begin{aligned}
l_\infty &= l^*(T_pR_p...T_1R_1(\Sigma^*)) \\
&\geq l^*(R_pT_{p-1}...T_1R_1(\Sigma^*)) \\
&\geq l^*(T_{p-1}R_{p-1}...T_1R_1(\Sigma^*)) \\
&\geq l^*(R_1(\Sigma^*)) \\
&= l_\infty.
\end{aligned}$$

This in particular implies that all the above are in fact equalities and hence $\forall i = 1, ..., k$, $T_i(\Sigma^*) = \Sigma^*$ and $R_i(\Sigma^*) = \Sigma^*$. Thus by the previous lemma, Σ^* is a stationary point $l^*(.)$. □

Definition 27. *$A, B \subset X$ are seperated if $(A \cap cl(B)) \cup (B \cap cl(A)) = \emptyset$.*

Note that if A and B are disjoint and closed, then A and B are seperated.

Definition 28. *X is connected if there doesn't exist $A, B \subset X$ such that A and B are seperated and $X = A \cup B$.*

Note that, in particular, X is connected if $X \neq A \cup B$, where A, B are disjoint and closed realted to X. Additionally, a compact set is said to be connected if it cannot be partitioned into two nonempty compact sets.

Lemma 22. *Every closed subset of a compact set is compact.*

Definition 29. *A point a is said to be an accumulation point of X if and only if $\forall \delta > 0$, $B_\delta(a)$ contains infinitely many points of X.*

Lemma 23. *A point a is an accumulation point for $E \subset X$ if and only if there is a sequence $x_n \in E \setminus \{a\}$ such that $x_n \to a$ as $n \to \infty$.*

Proof. If a is an accumulation point for E, then $(B_{\frac{1}{n}}(a) \cap E)$ contains infinitely many points. Thus let $x_n \in (B_{\frac{1}{n}}(a) \cap E) \setminus \{a\}$. Since $0 < d(x_n, a) < \frac{1}{n} \implies x_n \to a$ as $n \to \infty$.

Conversely suppose that $x_n \in E \setminus \{a\}$ and $x_n \to a$ as $n \to \infty$. Given $r > 0, \exists n_1,$ such that $d(x_{n_1}, a) < r \implies x_{n_1} \in B_r(a)$. We want to show that there are infinitely many $x_{n_k} \in B_r(a)$. One way to do this would be to pick $s < r$ such that $x_{n_2} \neq x_{n_1}, x_{n_2} \in B_s(a)$, but this is always possbile as we can define $r_1 = d(x_{n_1}, a)$ and then $s = \frac{r_1}{2}$. Continue in this fashion to find x_{n_3} and so on.

A more rigorous, but less intuitive proof is to assume that the neighborhood contains finitely many points, choose a distance smaller than the minimum and reach a contradiction but finding a ball with no elements in it. □

Lemma 24. *A point a is an accumulation point for $E \subset X$ if and only if $\forall r > 0, E \cap B_r(a) \setminus \{a\}$ is non-empty.*

Lemma 25. *Let*

$$A_\infty = \textit{ set of accumulation points of } \{\Sigma^r\}_{r \geq 0}$$

Then A_∞ is a compact subset of Θ_0.

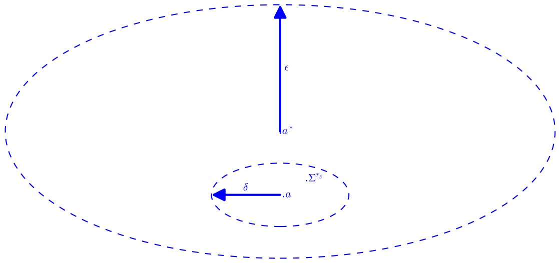


Figure 10: The plot shows $\Sigma^{r_\delta} \in B_\delta(a) \subset B_\epsilon(a^*)$.

Proof. Since A_∞ is a subset of a compact set, it suffices to show that A_∞ is closed, which is equivalent to showing that $A_\infty = cl(A_\infty)$. Let $a^* \in cl(A_\infty)$. Then for any $\epsilon > 0$, there exists $a \in A_\infty$ such that $a \in B_\epsilon(a^*)$. Similarly, since a is an accumulation point of Σ^r, $\forall \delta > 0, \exists r_\delta \in \mathbb{N}$ such that $\Sigma^{r_\delta} \in B_\delta(a)$. As $B_\epsilon(a^*)$ is open, we can pick δ small enough such that $B_\delta(a) \subset B_\epsilon(a^*)$, which implies that $\Sigma^{r_\delta} \in \subset B_\epsilon(a^*)$. Thus we have that $\forall \epsilon > 0, \exists \Sigma^{r_\delta} \in B_\epsilon(a^*)$ which implies that a^* is an accumulation point of $\{\Sigma^r\}_{r \geq 0}$, i.e. $a^* \in A_\infty$. See Figure 10 for an illustration. □

Lemma 26. *If A and B are disjoint compact sets, then $d(A, B) > 0$.*

Lemma 27. *Let*

$$A_\infty = \textit{ set of accumulation points of } \{\Sigma^r\}_{r \geq 0}$$

Then A_∞ is a connected subset of Θ_0.

Proof. Note that we already have that A_∞ is closed. Hence it contains all its limit points. Thus if $\Sigma^n \in A_\infty$, then $\forall \epsilon > 0, B_\epsilon(\Sigma^n) \cap \{\Sigma^r\}_{r \geq 0} \setminus \Sigma^n \neq \emptyset$, i.e. $\forall \epsilon > 0$, the ϵ-ball around Σ^n will contain a point from $\{\Sigma^r\}_{r \geq 0}$, say Σ^k distinct from Σ^n, i.e. $d(\Sigma^k, \Sigma^n) < \epsilon$. Equivalently, we can say that $\forall \epsilon > 0, B_\epsilon(\Sigma^n)$ will contain infinitely many Σ^r.

If A_∞ is not connected then there exists non-empty compact sets A, B such that $A_\infty = A \cup B$ and $A \cap B = \emptyset$. We will show that $B = \emptyset$, which is a contradiction and hence A_∞ must be connected. Let $B_\epsilon(A_\infty)$ be all $\Sigma \in \Theta_0 \implies l^*(\Sigma) \leq l^*(\Sigma^0)$ such that the distance of Σ from A_∞ is less than ϵ.

$$B_\epsilon(A_\infty) := \{\Sigma \in \Theta_0 : d(\Sigma, A_\infty) < \epsilon\}$$

where $d(\Sigma, A_\infty) = \inf_{\Sigma^r \in A_\infty} d(\Sigma, \Sigma^r)$. Thus if $\Sigma \in B_\epsilon(A_\infty)$, then there $\exists \Sigma^k \in A_\infty$ such that $d(\Sigma, \Sigma^k) < \epsilon$. Now let $\epsilon > 0$. Then we can find some $n_\epsilon \in \mathbb{N}$ such that for all $n \geq n_\epsilon, d(\Sigma^n, A_\infty) < \epsilon$. **I wasn't able to prove this directly but it makes sense that after some point all points will get arbitrarily close to some accumulation point. If this holds the rest is not that difficult.** First suppose $A_\infty = A \cup B$ where A, B are disjoint, compact sets. Then $d(A, B) > 0$. Set $\delta = \frac{d(A,B)}{2}$. As $S = T_p R_p ... T_1 R_1$ is continuous, where $\Sigma^{n+1} = S(\Sigma^n)$, S is uniformly continuous on Θ_0, which is compact. Thus $\forall \delta > 0, \exists \epsilon' > 0$ such that $a \in A_\infty$ and $\Sigma^n \in B_{\epsilon'}(a) \implies \Sigma^{n+1} \in B_\delta(a)$. Note that we can always pick ϵ' to be smaller than δ. Then for $n > n_\epsilon \implies n+1 > n_\epsilon \implies \Sigma^{n+1} \in B_\epsilon(A_\infty)$ and $\Sigma^n \in B_\epsilon(A) \implies \Sigma^{n+1} \in B_\delta(A)$. Also, $A \subset A_\infty$. Thus we have

$$\Sigma^{n+1} \in B_\delta(A) \cap B_\epsilon(A_\infty) = B_\epsilon(A)$$

As this hold for all $n+1 > n > n_\epsilon$, any point in B can have at most finitely many points near it and hence $B = \emptyset$. Thus A_∞ must be connected. □

Lemma 28. *Let*

$$A_\infty = \text{ set of accumulation points of } \{\Sigma^r\}_{r \geq 0}$$

be finite. Then $A_\infty = \{\Sigma^*\}$,i.e. $\exists \Sigma^*$ such that $\Sigma^r \to \Sigma^* \in A_\infty$.

Proof. By Bolzano-Weierstrass, $\exists \Sigma^*$ such that $\Sigma^{r_k} \to \Sigma^* \in A_\infty$. Hence A_∞ is non-empty. But A_∞ is connected and finite (by assumption) and hence must be a singleton, i.e. $\Sigma^r \to \Sigma^*$. □

Lemma 29. *Suppose*

$$l^*(\Sigma) = tr(\Sigma^{-1} S) + log|\Sigma|, \quad \Sigma \in \mathbb{P}_G.$$

has finitely many solutions on the same contour. Then $\Sigma^r \to \Sigma^*$.

Proof. This is true by Lemma 21 and Lemma 28. □

7.2 Kauermann's Dual Estimation

Finding $\hat{\Sigma}_{mle}$ is difficult in general. If $n < p$, S is singular and no mle estimate exists. Even if $n \geq p$, it is a highly non-convex problem. We just saw the proof of the ICF, which is a partial minimization problem, but due to non-convexity, proving convergence is difficult. Cox and Wermuth (1993) suggest changing the objective function to make it nicer. This is a very common approach. Kauermann (1996) suggest using the dual of the KullbackLeibler (KL) divergence to get a nicer form. Note that

$$l^*(\Sigma) = tr(\Sigma^{-1}S) + \log(\Sigma), \quad \Sigma \in \mathbb{P}_G$$

is not convex whereas

$$l^*(\Omega) = tr(\Omega S) - \log(\Omega), \quad \Omega \in \mathbb{P}_G$$

is convex.

Definition 30. *The KL divergence from $\mathcal{N}_0(\mu_0, \Sigma_0)$ to $\mathcal{N}_1(\mu_1, \Sigma_1)$ is*

$$\begin{aligned} D_{KL}(\mathcal{N}_0 || \mathcal{N}_1) &= \mathbb{E}_{\mathcal{N}_0}[\log \frac{f_{\mathcal{N}_0}(x)}{f_{\mathcal{N}_1}(x)}] \\ &= \int_{-\infty}^{\infty} [\log \frac{f_{\mathcal{N}_0}(x)}{f_{\mathcal{N}_1}(x)}] f_{\mathcal{N}_0}(x) dx \\ &= \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - p + \log \frac{|\Sigma_1|}{|\Sigma_0|}) \end{aligned}$$

where p is the dimension of the vector space.

In general it turns out that $\hat{\theta}_{mle} = \arg\min_{\theta_1 \in \Theta} D_{KL}(\mathbb{F}_n || \mathcal{N}(0, \theta_1))$. Note that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} -\log f_{\theta_1}(X_i) \to -\mathbb{E}[\log f_{\theta_1}(X_1)] \quad a.s.$$

by the strong law of large numbers. Note that if we take the derivative with respect to θ_1, then we get the likelihood equations. Now suppose $\theta_1 = \Sigma$.

Then

$$\begin{aligned}
-\mathbb{E}[\log f_{\Sigma}(X_1)] &= \mathbb{E}[\frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}x^T\Sigma^{-1}x] \\
&= \frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}\mathbb{E}[x^T\Sigma^{-1}x] \\
&= \frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}\mathbb{E}[tr(x^T\Sigma^{-1}x)] \\
&= \frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}\mathbb{E}[tr(\Sigma^{-1}xx^T)] \\
&= \frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}tr(\mathbb{E}[\Sigma^{-1}xx^T]) \\
&= \frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}tr(I_p) \\
&= \frac{p}{2}\log(2\pi) + \log|\Sigma| + \frac{1}{2}p.
\end{aligned}$$

Note that this also implies we can use the negative log-likelihood as a proxy for $\mathbb{E}[\log f_{\theta_1}(X_1)]$. Now suppose θ_0 is the true parameter. Then,

$$\begin{aligned}
\mathbb{E}_{\theta_0}[\log f_{\theta_0}(X_1) - \log f_{\theta_1}(X_1)] &= \mathbb{E}_{\theta_0}[\log \frac{f_{\theta_0}(X_1)}{f_{\theta_1}(X_1)}] \\
&= D_{KL}(f_{\theta_0}||f_{\theta_1}).
\end{aligned}$$

More importantly,

$$\begin{aligned}
-D_{KL}(f_{\theta_0}||f_{\theta_1}) &= -\mathbb{E}_{\theta_0}[\log f_{\theta_0}(X_1) - \log f_{\theta_1}(X_1)] \quad = \mathbb{E}_{\theta_0}[\log \frac{f_{\theta_1}(X_1)}{f_{\theta_0}(X_1)}] \\
&\leq \log \mathbb{E}_{\theta_0}[\frac{f_{\theta_1}(X_1)}{f_{\theta_0}(X_1)}] \\
&= \log \int \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} f_{\theta_0}(X_1)dx \\
&= \log \int f_{\theta_0}(X_1)dx \\
&= \log 1 \\
&= 0 \implies D_{KL}(f_{\theta_0}||f_{\theta_1}) \geq 0.
\end{aligned}$$

In using the negative log-likelihood risk function, we define the empirical risk, $\hat{R}_n := -\frac{1}{n}\sum_{i=1}^{n} \log f_{\theta_0}(X_1)$ and select the distribution that minimizes empirical risk, i.e. $\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{R}_n$. Now define the risk, $R(\theta) := -\mathbb{E}_\theta[\log f_{\theta_0}(X_1)]$. Then the excess risk is equal to,

$$\begin{aligned} R(\theta) - R(\theta_1) &= -\mathbb{E}_\theta[\log f_\theta(X_1)] + \mathbb{E}_\theta[\log f_{\theta_1}(X_1)] \\ &= D_{KL}(f_\theta || f_{\theta_1}) \end{aligned}$$

which is minimized at $\theta_1 = \theta$. Now recall that as

$$\begin{aligned} &\sqrt{n}(\mathbb{F}_n(t) - F(t)) \xrightarrow{d} \mathcal{N}(0, F(t)(1 - F(t))) \\ &\implies \mathbb{F}_n(t) \dot{\sim} \mathcal{N}(\frac{F(t)}{\sqrt{n}}, \frac{F(t)(1 - F(t))}{n}), \end{aligned}$$

where $F \sim \mathcal{N}(0, \Sigma_1)$. Then we can suppose that $\mathbb{F}_n \dot{\sim} \mathcal{N}(0, S)$ where $\mathbb{E}[S] = \Sigma_1$, that is, $\mu_1 = 0$ and $(\mu_0, \Sigma_0) = (0, S)$. Then

$$D_{KL}(\mathcal{N}_0 || \mathcal{N}_1) = \frac{1}{2}(tr(\Sigma_1^{-1}S) - p + \log\frac{|\Sigma_1|}{|S|})$$

Then $\hat{\Sigma}_{1,mle}$ can be found by minimizing the KL divergence with respect to Σ_1.

$$\begin{aligned} \hat{\Sigma}_{1,mle} &= \arg\min_{\Sigma_1 \in \mathbb{P}_G} \frac{1}{2}(tr(\Sigma_1^{-1}S) - p + \log\frac{|\Sigma_1|}{|S|}) \\ &= \arg\min_{\Sigma_1 \in \mathbb{P}_G} \frac{1}{2}(tr(\Sigma_1^{-1}S) + \log|\Sigma_1|) \end{aligned}$$

Since the Kullback-Leibler information is not symmetrical in its arguments, we get a different minimization problem if the observed and unknown parameters are exchanged. This yields the dual maximum likelihood estimator, $\hat{\Sigma}_{1,dual}$, obtained by

$$\begin{aligned} \hat{\Sigma}_{1,dual} &= \arg\min_{\Sigma_1 \in \mathbb{P}_G} \frac{1}{2}(tr(\Sigma_1 S^{-1}) - p + \log\frac{|S|}{|\Sigma_1|}) \\ &= \arg\min_{\Sigma_1 \in \mathbb{P}_G} \frac{1}{2}(tr(\Sigma_1 S^{-1}) - \log|\Sigma_1|) \end{aligned}$$

which is a convex problem. However, to solve this problem we must be able to find S^{-1} which means that $n > p$. This is one reason to use the Bayesian framework in the case when $n < p$.

7.3 The Inverse G-Wishart Distribution

Definition 31. *Recall that an **exponential family** is defined to be one where we can factorize the density as follows:*

$$f(x|\theta) = h(x)e^{\eta(\theta)^T T(x) - A(\theta)}$$

or, equivalently

$$f(x|\theta) = h(x)g(\theta)e^{\eta(\theta)^T T(x)}.$$

*This is said to be in **canonical form** if $\forall i, \eta_i(\theta) = \theta_i$. If the dimension of $\theta = (\theta_1, ..., \theta_d)$ is less than the dimension of $\eta(\theta) = (\eta_1(\theta), ..., \eta_s(\theta))$, then this is called a **curved** exponential family.*

The simplest example of a curved exponential family is the $\mathcal{N}(\theta, \theta^2)$.

$$\begin{aligned} f(x|\theta) &= \frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{x^2}{2\theta^2} + \frac{x}{\theta} - \frac{1}{2}} \\ &= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi\theta^2}} e^{(-\frac{1}{2\theta^2}, \frac{1}{\theta})\begin{pmatrix} x^2 \\ x \end{pmatrix}} \end{aligned}$$

where the dimension of θ is 1, which is smaller than the dimension of $\eta(\theta) = (-\frac{1}{2\theta^2}, \frac{1}{\theta})$, which is equal to 2. Recall that S is sufficient for Σ and hence consider

$$\begin{aligned} L(\Sigma) &= \prod_{i=1}^{n} f(x^i|\Sigma) \\ &= e^{-\frac{n}{2}tr(\Sigma^{-1}S) - \frac{n}{2}\log|\Sigma|}, \quad \Sigma \in \mathbb{P}_G \end{aligned}$$

Note that this is not the kernel of a natural exponential family, but a curved exponential family due to the constraint that $\Sigma \in \mathbb{P}_G$. Silva and Ghahramamni (2009) introduce the following inverse $G-Wishart$ prior ($\mathcal{GIW}(\mathcal{U}, \delta)$) with density, π,

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

Then as $tr(A + B) = tr(A) + tr(B)$, we have that,

$$\begin{aligned} \pi(\Sigma|Data) &= \pi(\Sigma|S) \\ &\propto L(\Sigma)\pi_{U,\delta}(\Sigma) \\ &\propto e^{-\frac{1}{2}tr(\Sigma^{-1}(nS+U)) - \frac{\delta+n}{2}\log|\Sigma|} \\ \implies (\Sigma|Data) &\sim \mathcal{GIW}(nS + U, n + \delta). \end{aligned}$$

Note that for any prior to be useful we must either be able to calculate the posterior expectation easily or we must be able to sample from the posterior in a computationally feasilbe way. To that extent consider the modified Cholesky decompostion of $\Sigma = LDL^T$ where L is a lower triangular matrix with 1's on the diagonal and D is diagonal. Recall that this is a bijection as it is a unique decomposition, i.e. given any Σ such that $(i,j) \notin E \implies \Sigma_{ij} = 0$ we can find the corresponding (L, D) such that $(i,j) \notin E \implies L_{ij} = 0$ and given any (L, D) we can find the corresponding Σ. In short,

$$\{\Sigma_{ij}\}_{i \geq j, (i,j) \in E} \mapsto (\{L_{ij}\}_{i \geq j, (i,j) \in E}, D)$$

is a bijection from $\mathbb{P}_G$ to $\mathbb{R}^{|E|} \times \mathbb{R}^p_+$. Also recall that in Section 6.10 we showed that the Jacobian of this transformation, $|J| = \prod_{i=1}^p D_{jj}^{n_j}$ where $n_j = |\{i | i > j, (i, j \in E)\}|$. Thus,

$$\begin{aligned}
\pi_{U,\delta}(L, D) &= \pi_{U,\delta}(\Sigma(L, D))|J| \\
&= e^{-\frac{1}{2} tr((\Sigma(L,D))^{-1}U) - \frac{\delta}{2} \log|\Sigma(L,D)|} \prod_{j=1}^p D_{jj}^{n_j} \\
&= e^{-\frac{1}{2} tr((L^T)^{-1}D^{-1}L^{-1}U) - \frac{\delta}{2} \log|LDL^T|} e^{\sum_{j=1}^p \log D_{jj}^{n_j}} \\
&= e^{-\frac{1}{2} tr(D^{-1}L^{-1}U(L^T)^{-1}) - \frac{\delta}{2} \log|L||D||L^T|} e^{\sum_{j=1}^p \log D_{jj}^{n_j}} \\
&= e^{-\frac{1}{2} \sum_{i=1}^p (D^{-1}L^{-1}U(L^T)^{-1})_{ii} - \frac{\delta}{2} \log|D| + \sum_{j=1}^p \log D_{jj}^{n_j}} \\
&= e^{-\frac{1}{2} \sum_{i=1}^p (D^{-1})_{ii}(L^{-1}U(L^T)^{-1})_{ii} - \sum_{i=1}^p \frac{\delta}{2} \log D_{jj} + \sum_{j=1}^p n_j \log D_{jj}} \\
&= e^{-\sum_{i=1}^p \frac{(L^{-1}U(L^T)^{-1})_{ii}}{2D_{ii}} - \sum_{j=1}^p (\frac{\delta}{2} - n_j) \log D_{jj}} \\
&= \prod_{i=1}^p e^{-\frac{(L^{-1}U(L^T)^{-1})_{ii}}{2D_{ii}}} \prod_{j=1}^p D_{jj}^{-\frac{\delta - 2n_j}{2}} \\
&= \prod_{i=1}^p \left\{ e^{-\frac{(L^{-1}U(L^T)^{-1})_{ii}}{2D_{ii}}} D_{ii}^{-\frac{\delta - 2n_i}{2}} \right\}
\end{aligned}$$

Definition 32. *If $X \sim \Gamma(\alpha, \beta)$ with pdf $f_X(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha-1}}$, then $\frac{1}{X} \sim \mathcal{I}\Gamma(\alpha, \frac{1}{\beta})$ with pdf $f_{\frac{1}{X}}(x) = \frac{x^{-\alpha-1}e^{-\frac{\beta}{x}}\beta^\alpha}{\Gamma(\alpha)}$ where $\mathcal{I}\Gamma(\alpha, \frac{1}{\beta})$ denotes the **inverse gamma distribution**.*

We immediately notice that $\pi_{U,\delta}(D|L)$ would resembles indepedent $\mathcal{I}\Gamma$ densities, but the problem is that $(i,j) \notin E \implies L_{ij}$ is a dependent entry in the sense that we can write all such L_{ij} as a function of the independent entries L_{kl} such that $(k,l) \in E$ and D_{kk}. Thus we need to find a transformation that will free us of these dependencies. Consider the transformation,

$$(\{L_{ij}\}_{i \geq j, (i,j) \in E}, D) \mapsto (\{L_{ij}^{-1}\}_{i \geq j, (i,j) \in E}, D).$$

To see that J will be triangular matrix with 1's on the diagonal, note that if we let $T = L^{-1}$, then T is also lower triangular and $\frac{d}{dT_{ij}}(T^{-1})_{kl} = (T^{-1})_{ik}(T^{-1})_{lj}$. Thus for the diagonal elements of J, $\frac{d}{dT_{ij}}(T^{-1})_{ij} = (T^{-1})_{ii}(T^{-1})_{jj} = 1$ as $L_{ii} = 1$ and for $k > i$, $\frac{d}{dT_{ij}}(T^{-1})_{kl} = (T^{-1})_{ik}(T^{-1})_{lj} = 0$ as $(T^{-1})_{ik} = 0$. Thus, in the simple case when $p = 3$, the Jacobian matrix would look like

	d_{11}	l_{21}	d_{22}	l_{31}	l_{32}	d_{33}
d_{11}	1	0	0	0	0	0
t_{21}	0	1	0	0	0	0
d_{22}	0	0	1	0	0	0
t_{31}	0	$*$	0	1	0	0
t_{32}	0	$*$	0	$*$	1	0
d_{33}	0	0	0	0	0	1

which implies that $|J| = 1$. Thus

$$\begin{aligned}
\int_{\mathbb{R}^{|E|}} \int_{\mathbb{R}^p_+} \pi_{U,\delta}(D,L) dD dL &= \int_{\mathbb{R}^{|E|}} \int_{\mathbb{R}^p} \pi_{U,\delta}(T,L) |J| dT dL \\
&= \int_{\mathbb{R}^{|E|}} \int_{\mathbb{R}^p_+} \pi_{U,\delta}(T,L) dT dL \\
&= \int_{\mathbb{R}^{|E|}} \int_{\mathbb{R}^p_+} \prod_{i=1}^{p} \left\{ e^{-\frac{(TUT^T)_{ii}}{2D_{ii}}} D_{ii}^{-\frac{\delta - 2n_i}{2}} \right\} dT dL
\end{aligned}$$

Now let $U = \Phi \Lambda \Phi^T$ be the modified cholesky decompostion for U and let

$\tilde{T} = T\Phi$. Then

$$\begin{aligned}
(TUT^T)_{ii} &= (T\Phi\Lambda\Phi^T T^T)_{ii} \\
&= (T\Phi\Lambda\Phi^T T^T)_{ii} \\
&= (\tilde{T}\Lambda\tilde{T}^T)_{ii} \\
&= \tilde{T}_{i.}\Lambda\tilde{T}_{i.}^T \\
&= (\tilde{t}_{i1}, \dots, \tilde{t}_{i,i-1}, 1, 0, \dots, 0) \begin{pmatrix} \lambda_{11} & 0 & 0 & \dots & 0 \\ 0 & \lambda_{22} & 0 & \dots & 0 \\ 0 & 0 & \lambda_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \lambda_{pp} \end{pmatrix} \begin{pmatrix} \tilde{t}_{i1} \\ \vdots \\ \tilde{t}_{i,i-1} \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\
&= \sum_{j=1}^{i-1} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii} \\
&= \sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \sum_{j<i,(i,i)\notin E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii}
\end{aligned}$$

Note that if U is positive definite, then $\forall j, \lambda_{jj} > 0$ which gives us

$$\textbf{Condition 1: } (TUT^T)_{ii} \geq \sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii}$$

Aditiionally, as $D_{ii} > 0$,

$$\begin{aligned}
-(TUT^T)_{ii} &\leq - \sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii} \\
\implies \frac{-(TUT^T)_{ii}}{2D_{ii}} &\leq \frac{-\sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii}}{2D_{ii}} \\
\implies e^{-\frac{(TUT^T)_{ii}}{2D_{ii}}} &\leq e^{\frac{-\sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii}}{2D_{ii}}}
\end{aligned}$$

Note also that the Jabobian for the transformation $T \mapsto \tilde{T}$ is independent of $\tilde{T}$ as $T = \tilde{T}\Phi^{-1} = \tilde{T}\Phi^T$ which implies that

$$\frac{d}{d\tilde{T}}T = \frac{d}{d\tilde{T}}\tilde{T}\Phi^T = \Phi$$

. Thus,

$$\begin{aligned}
&\int_{\mathbb{R}^{|E|}}\int_{\mathbb{R}_+^p}\prod_{i=1}^p e^{-\frac{(TUT^T)_{ii}}{2D_{ii}}} D_{ii}^{-\frac{\delta-2n_i}{2}}dDdT\\
&\leq \int_{\mathbb{R}^{|E|}}\int_{\mathbb{R}_+^p}\prod_{i=1}^p e^{\frac{-\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii}}{2D_{ii}}} D_{ii}^{-\frac{\delta-2n_i}{2}}|J|dDd\tilde{T}\\
&= |J|\int_{\mathbb{R}^{|E|}}\prod_{i=1}^p\int_{\mathbb{R}_+} e^{\frac{-\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii}}{2D_{ii}}} D_{ii}^{-\frac{\delta-2n_i}{2}}dDd\tilde{T}\\
&= |J|\int_{\mathbb{R}^{|E|}}\prod_{i=1}^p\int_{\mathbb{R}_+} e^{-\frac{1}{2D_{ii}}(\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii})} D_{ii}^{-\frac{\delta-2n_i}{2}}dDd\tilde{T}\\
&= |J|\int_{\mathbb{R}^{|E|}}\prod_{i=1}^p\frac{\Gamma(-\frac{\delta-2n_i}{2}-1)}{\frac{1}{2}(\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}}d\tilde{T}
\end{aligned}$$

provided $-\frac{\delta-2n_i}{2}-1>0 \iff \frac{2n_i-\delta}{2}-1>0$

$$\textbf{Condition 2: } \frac{2n_i-\delta}{2}-1>0 \quad \forall 1\leq i\leq p.$$

Now,

$$\begin{aligned}
&\int_{\mathbb{R}^{|E|}}\int_{\mathbb{R}_+^p}\pi_{U,\delta}(D,L)dDdL\\
&= \int_{\mathbb{R}^{|E|}}\int_{\mathbb{R}_+^p}\prod_{i=1}^p e^{-\frac{(TUT^T)_{ii}}{2D_{ii}}} D_{ii}^{-\frac{\delta-2n_i}{2}}dDdT\\
&= |J|\int_{\mathbb{R}^{|E|}}\prod_{i=1}^p\frac{\Gamma(-\frac{\delta-2n_i}{2}-1)}{\frac{1}{2}(\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}}d\tilde{T}\\
&= |J|\int_{\mathbb{R}^{|E|}}\prod_{i=1}^p\Gamma(-\frac{\delta-2n_i}{2}-1)\frac{1}{\frac{1}{2}(\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}}d\tilde{T}
\end{aligned}$$

As $\Gamma(-\frac{\delta-2n_i}{2}-1)$ is a constant for each i, we only focus on

$$\int_{\mathbb{R}^{|E|}}\prod_{i=1}^p\frac{1}{\frac{1}{2}(\sum_{j<i,(i,j)\in E}\lambda_{jj}\tilde{t}_{ij}^2+\lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}}d\tilde{T}.$$

For $2 \leq i \leq p$, let

$$x^i := \begin{pmatrix} \sqrt{\lambda_{1,1}}\tilde{t}_{i,1} \\ \vdots \\ \sqrt{\lambda_{i-1,i-1}}\tilde{t}_{i,i-1} \end{pmatrix}$$

which implies that $\|x^i\|^2 = (x^i)^T x^i = \sum_{j=1}^{i-1} \lambda_{jj}\tilde{t}_{ij}^2$ and let $v_i = |\{j : j < i, (i,j) \in E\}|$. Also note that if L is the lower triangular matrix such that $L_{.i} = (1, x^i)^T$, then $L = \Lambda^{\frac{1}{2}}T$ which implies $T = \Lambda^{-\frac{1}{2}}L$ and $\frac{d}{dL}T = \Lambda^{-\frac{T}{2}}$. Thus the Jacobian of the transformation $T \mapsto L$ is independent of L. Therefore,

$$\begin{aligned}
&|J| \int_{\mathbb{R}^{|E|}} \prod_{i=1}^{p} \Gamma(-\frac{\delta - 2n_i}{2} - 1) \frac{1}{\frac{1}{2}(\sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}} d\tilde{T} \\
\propto &\int_{\mathbb{R}^{|E|}} \prod_{i=1}^{p} \frac{1}{\frac{1}{2}(\sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}} d\tilde{T} \\
= &\frac{1}{\frac{1}{2}\lambda_{11}^{-\frac{\delta-2n_1}{2}-1}} \int_{\mathbb{R}^{|E|}} \prod_{i=2}^{p} \frac{1}{\frac{1}{2}(\sum_{j<i,(i,j)\in E} \lambda_{jj}\tilde{t}_{ij}^2 + \lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}} d\tilde{T} \\
\propto &\prod_{i=2}^{p} \int_{\mathbb{R}^{v_i}} \frac{1}{\frac{1}{2}(\|x^i\|^2 + \lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}} dx^i \\
\propto &\prod_{i=2}^{p} \int_{\mathbb{R}^{v_i}} \frac{1}{(\|x^i\|^2 + \lambda_{ii})^{-\frac{\delta-2n_i}{2}-1}} dx^i \\
= &\prod_{i=2}^{p} \int_{\mathbb{R}^{v_i}} \frac{1}{\lambda_{ii}^{-\frac{\delta-2n_i}{2}-1}} \frac{1}{(1 + \frac{\|x^i\|^2}{\lambda_{ii}})^{-\frac{\delta-2n_i}{2}-1}} dx^i \\
\propto &\prod_{i=2}^{p} \int_{\mathbb{R}^{v_i}} \frac{1}{(1 + \frac{\|x^i\|^2}{\lambda_{ii}})^{-\frac{\delta-2n_i}{2}-1}} dx^i
\end{aligned}$$

which is finite if and only if $-\frac{\delta-2n_i}{2} - 1 > \frac{v_i}{2}$.

$$\textbf{Condition 3: } \frac{2n_i - \delta}{2} - 1 > \frac{v_i}{2} \quad \forall 2 \leq i \leq p$$

by Lemma 30.

Lemma 30.

$$\int_{\mathbb{R}} \frac{1}{(1+x^2)^b} dx < \infty \iff b > \frac{1}{2}.$$

Proof. Proof needed. □

Thus $\pi_{U,\delta}(\sigma)$ is a proper prior, i.e. it is a probablity density, if and only if the following hold:

1. U is positive definite.
2. $2n_1 - \delta > 2$.
3. $2n_i - \delta > v_i + 2 \geq 2$ for $j = 2, ..., p$.

7.4 Gibbs Sampling for $\mathcal{GIW}(U, \delta)$

Recall that in the case of the the concentration matrix, we could use a Gibbs-sampler in deriving the maximum likelihood estimator. Similarly, in the case of a covariance graph model, we can develop a Gibbs-sampling procedure to find the maximum likelihood estimator.

$$\pi_{U,\delta}(\Sigma) = e^{-\frac{1}{2}tr(\Sigma^{-1}U) - \frac{\delta}{2}\log|\Sigma|}$$

and consider the following partiotion of Σ,

$$\begin{aligned} \Sigma &= \begin{pmatrix} \Sigma_{11} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11} & \Sigma_{-1,1}^T \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix} \end{aligned}$$

Let $\gamma_1 = \Sigma_{11} - \Sigma_{-1,1}^T \Sigma_{-1,-1}^{-1} \Sigma_{-1,1}$. Then

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\gamma_1} & -\frac{1}{\gamma_1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1} \\ -\frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1} & \Sigma_{-1,-1}^{-1} + \frac{1}{\gamma_1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1} \end{pmatrix}$$

Note that if

$$\begin{aligned} U &= \begin{pmatrix} U_{11} & U_{1,-1} \\ U_{-1,1} & U_{-1,-1} \end{pmatrix} \\ &= \begin{pmatrix} U_{11} & U_{-1,1}^T \\ U_{-1,1} & U_{-1,-1} \end{pmatrix} \end{aligned}$$

then

$$\begin{aligned}
tr(\Sigma^{-1}U) &= tr(\frac{1}{\gamma_1}U_{11} - \frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1}) \\
&\quad + tr(-\frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}U^T_{-1,1} + \Sigma^{-1}_{-1,-1} + \frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}) \\
&= tr(\frac{1}{\gamma_1}U_{11}) - tr(\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1}) - tr(\frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}U^T_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1} + \frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}) \\
&= tr(\frac{1}{\gamma_1}U_{11}) - tr(\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1}) - tr(\frac{1}{\gamma_1}U^T_{-1,1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1} + \frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}) \\
&= tr(\frac{1}{\gamma_1}U_{11}) - 2tr(\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1}) + tr(\frac{1}{\gamma_1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}) \\
&= tr(\frac{1}{\gamma_1}U_{11}) - 2tr(\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1}) \\
&\quad + tr(\Sigma^{-1}_{-1,-1}) + tr(\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}) \\
&= tr(\frac{1}{\gamma_1}U_{11}) + tr(\Sigma^{-1}_{-1,-1}) \\
&\quad + (\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1}) - 2(\frac{1}{\gamma_1}\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1}) \\
&= tr(\frac{1}{\gamma_1}U_{11}) + tr(\Sigma^{-1}_{-1,-1}) \\
&\quad + \frac{1}{\gamma_1}(\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,-1}\Sigma^{-1}_{-1,-1}\Sigma_{-1,1} - 2\Sigma^T_{-1,1}\Sigma^{-1}_{-1,-1}U_{-1,1})
\end{aligned}$$

and

$$\begin{aligned}
\log|\Sigma| &= \log|\gamma_1||\Sigma_{-1,-1}| \\
&= \log\gamma_1|\Sigma_{-1,-1}| \\
&= \log\gamma_1 + \log|\Sigma_{-1,-1}|
\end{aligned}$$

Now $\Sigma \mapsto (\Sigma_{1,-1}, \Sigma_{-1,-1}, \gamma_1)$ is a bijection. To understand what the Jacobian might look like look, suppose $p = 3$. Then note that

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{11}} \gamma_1 &= 1 \\
\frac{\partial}{\partial \Sigma_{-1,1}} \gamma_1 &= 2\Sigma_{-1,-1}^{-1}\Sigma_{-1,1} \\
\frac{\partial}{\partial \Sigma_{-1,-1}} \gamma_1 &= \Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Sigma_{-1,1}^{T}\Sigma_{-1,-1}^{-1} \\
\frac{\partial}{\partial \Sigma_{11}} \Sigma_{-1,1} &= 0 \\
\frac{\partial}{\partial \Sigma_{-1,1}} \Sigma_{-1,1} &= 1 \\
\frac{\partial}{\partial \Sigma_{-1,-1}} \Sigma_{-1,1} &= 0 \\
\frac{\partial}{\partial \Sigma_{11}} \Sigma_{-1,-1} &= 0 \\
\frac{\partial}{\partial \Sigma_{-1,1}} \Sigma_{-1,-1} &= 0 \\
\frac{\partial}{\partial \Sigma_{-1,-1}} \Sigma_{-1,-1} &= 1
\end{aligned}$$

which implies that

	γ_1	Σ_{12}	Σ_{13}	Σ_{22}	Σ_{23}	Σ_{33}
Σ_{11}	1	$*$	$*$	$*$	$*$	$*$
Σ_{12}	0	1	0	0	0	0
Σ_{13}	0	0	1	0	0	0
Σ_{22}	0	0	0	1	0	0
Σ_{23}	0	0	0	0	1	0
Σ_{33}	0	0	0	0	0	1

Hence the Jacobian, $|J| = 1$. Thus

$$\begin{aligned}
&\pi_{U,\delta}(\Sigma_{-1,1}, \Sigma_{-1,-1}, \gamma_1) \\
&= e^{-\frac{1}{2}tr(\Sigma^{-1}(\Sigma_{-1,1},\Sigma_{-1,-1},\gamma_1)U)-\frac{\delta}{2}\log|\Sigma(\Sigma_{-1,1},\Sigma_{-1,-1},\gamma_1)|}|J| \\
&= e^{-\frac{1}{2}(\frac{1}{\gamma_1}U_{11}+tr(\Sigma_{-1,-1}^{-1})+\frac{1}{\gamma_1}(\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}-2\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,1}))} \\
&\quad \times e^{\frac{\delta}{2}(\log\gamma_1+\log|\Sigma_{-1,-1}|)}
\end{aligned}$$

Recall that $X \sim \mathcal{N}(\mu, \Sigma)$ if

$$\begin{aligned}
f_X(x) &\propto e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \\
&= e^{-\frac{1}{2}(x^T\Sigma^{-1}x-\mu^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu+\mu^T\Sigma^{-1}\mu)}
\end{aligned}$$

and note that

$$\begin{aligned}
&\pi_{U,\delta}(\Sigma_{-1,1}|\Sigma_{-1,-1}, \gamma_1) \\
&\propto e^{-\frac{1}{2\gamma_1}(\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}-2\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,1})} \\
&= e^{-\frac{1}{2}(\Sigma_{-1,1}^T\frac{\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}}{\gamma_1}\Sigma_{-1,1}-2\Sigma_{-1,1}^T\frac{\Sigma_{-1,-1}^{-1}}{\gamma_1}}U_{-1,1}) \\
&= e^{-\frac{1}{2}(\Sigma_{-1,1}^T\frac{\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}}{\gamma_1}\Sigma_{-1,1}-2\Sigma_{-1,1}^T\frac{\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}}{\gamma_1}\Sigma_{-1,-1}U_{-1,-1}^{-1}U_{-1,1})} \\
&\propto e^{-\frac{1}{2}(\Sigma_{1,-1}-\Sigma_{-1,-1}U_{-1,-1}^{-1}U_{-1,1})^T(\frac{\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}}{\gamma_1})(\Sigma_{1,-1}-\Sigma_{-1,-1}U_{-1,-1}^{-1}U_{-1,1})}
\end{aligned}$$

which implies that $(\Sigma_{1,-1}|\Sigma_{-1,-1}, \gamma_1) \sim \mathcal{N}(\Sigma_{-1,-1}U_{-1,-1}^{-1}U_{-1,1}, \gamma_1\Sigma_{-1,-1}U_{-1,-1}^{-1}\Sigma_{-1,-1})$. Now note that

$$\begin{aligned}
&\pi_{U,\delta}(\gamma_1|\Sigma_{-1,1}, \Sigma_{-1,-1}) \\
&= \gamma_1^{\frac{\delta}{2}}e^{-\frac{1}{2\gamma_1}(U_{11}+\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}-2\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,1})}
\end{aligned}$$

which is $\mathcal{I}\Gamma(\frac{\delta}{2}-1, \frac{1}{2}(U_{11}+\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}-2\Sigma_{-1,1}^T\Sigma_{-1,-1}^{-1}U_{-1,1}))$.

Theorem 33. *Let*

$$\Sigma = \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,-i} \\ \Sigma_{-i,i} & \Sigma_{-i,-i} \end{pmatrix}$$

and

$$\gamma_i = \Sigma_{i,i} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}$$

which implies that $\Sigma_{i,i} = \gamma_i + \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}$.

For $i = 1, ..., p$,

1. $\Sigma_{i,-i}|(\gamma_i, \Sigma_{-i,-i}) \sim \mathcal{N}(\Sigma_{-i,-i}U_{-i,-i}^{-1}U_{-i,i}, \gamma_1\Sigma_{-i,-i}U_{-i,-i}^{-1}\Sigma_{-i,-i})$

2. $\gamma_i|(\Sigma_{i,-i}, \Sigma_{-i,-i}) \sim \mathcal{I}\Gamma(\frac{\delta}{2}-1, \frac{1}{2}(U_{ii}+\Sigma_{-i,i}^T\Sigma_{-i,-i}^{-1}U_{-i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}-2\Sigma_{-i,i}^T\Sigma_{-i,-i}^{-1}U_{-i,i}))$

Then the stationary distribution of Σ *is* $\mathcal{GIW}(U, \delta)$*.*

In a standard Gibbs-Sampling algorithm to generate from (X, Y) we usually need the conditional distributions $X|Y$ and $Y|X$. Thus to generate from Σ the standard approach would be to generate from $\Sigma_{11}|(\Sigma_{1,-1}, \Sigma_{-1,-1})$, $\Sigma_{1,-1}|(\Sigma_{1,1}, \Sigma_{-1,-1})$ and $\Sigma_{-1,-1}|(\Sigma_{1,-1}, \Sigma_{1,1})$. This is not what we are doing in the above algorithm where the partition is changing at every step. However, as Asmussen and Glynn (2011) show, the chain still converges to the correct stationary distribution so long as the Markov Chain has positive transition probablity everywhere, which is true in this case. Essentially we have an irreducible and aperiodic markov chain, which guarantees that the markoc chain is ergodic, i.e.,

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n} f(X_i) &\overset{a.s.}{\to} \mathbb{E}_\pi[f(X)] \\ &= \int f(x)\pi(x)dx\end{aligned}$$

An ergodic markov chain, X_i converges to it unique stationary distribution, π.

7.5 Homogenous Graph

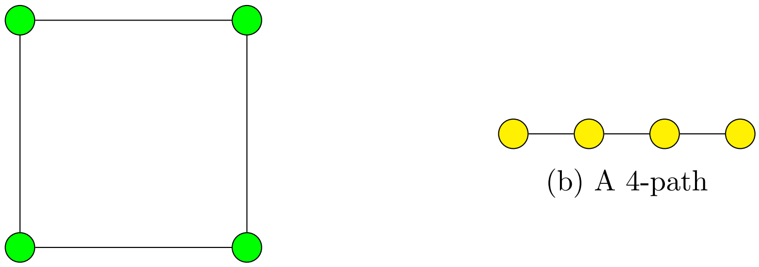
If the graph corresponding to the covariance matrix is homogenous, then

1. we can compute the maximum likelihood estimator,$\hat{\Sigma}_{mle}$ in closed form and

2. we can sample from the $\mathcal{GIW}$ density directly, without using the expensive Markov Chain mentioned in the previois section

Definition 34. *A graph* $G = (V, E)$ *is homogenous if it does not have a* $4-cycle$*, as shown in Figure 11a, or a* $4-path$*, as shown in Figure 11b, as an induced subgraph.*

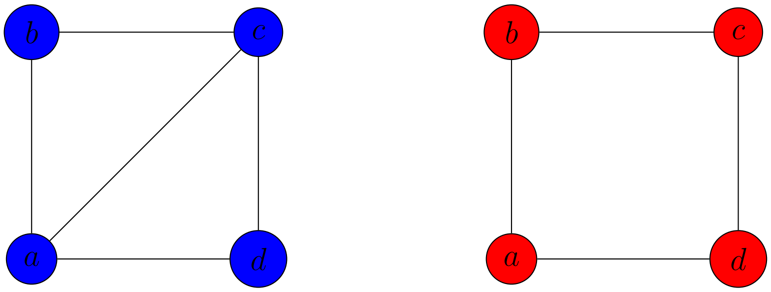
From a homogenous graph we construct a directed graph as described below.



(b) A 4-path

(a) A 4-cycle

Figure 11: A 4-cycle and a 4-path



(a) A decomposable graph (b) A non-decomposable graph

Figure 12: A decomposable and a non-decomposable graph

Lemma 31. *If $G = (V, U)$ is decomposable and $(u, v) \in E$, then either*

$$\{w : w = u \text{ or } (u, w) \in E\} \subset \{w : w = v \text{ or } (v, w) \in E\},$$

or

$$\{w : w = v \text{ or } (v, w) \in E\} \subset \{w : w = u \text{ or } (u, w) \in E\}.$$

Hence either we have that u or the neighbors of u are contained in v or the neighbors of v, or vice versa. In Figure 12a, we have that $E = \{(a, b), (b, c), (a, c), (c, d), (d, a)\}$. Also note that if $nb(u) := \{\text{neighbors of u}, u \in E\}$ and $N(u) := \{u\} \cup nb(u)$, then

$$\begin{aligned} nb(a) = \{b, c, d\} &\implies N(a) = \{a, b, c, d\} \\ nb(b) = \{a, c\} &\implies N(b) = \{a, b, c\} \\ nb(c) = \{b, a, d\} &\implies N(c) = \{a, b, c, d\} \\ nb(d) = \{a, c\} &\implies N(d) = \{a, c, d\} \end{aligned}$$

which implies that

$$\begin{aligned}N(b) &\subset N(a)\\ N(b) &\subset N(c)\\ N(a) &= N(c)\\ N(d) &\subset N(c)\\ N(d) &\subset N(a)\end{aligned}$$

However, in Figure 12b we have a non-decomposable graph such that $E = \{(a, b), (b, c), (c, d), (d, a)\}$ and

$$\begin{aligned}nb(a) = \{b, d\} &\implies N(a) = \{a, b, d\}\\ nb(b) = \{a, c\} &\implies N(b) = \{a, b, c\}\\ nb(c) = \{b, d\} &\implies N(c) = \{b, c, d\}\\ nb(d) = \{a, c\} &\implies N(d) = \{a, c, d\}.\end{aligned}$$

Here, $N(a) \not\subset N(b)$ and $N(b) \not\subset N(a)$.

Definition 35. *Let $G = (V, E)$ be a decomposable graph. Let $v \in V$.*

$$\bar{v} \coloneqq \{w : N(v) = N(w)\}.$$

Then

$$\bar{V} \coloneqq \{\bar{v} : v \in V\}$$

and

$$\bar{E} \coloneqq \{(\bar{u}, \bar{v}) : \bar{u}, \bar{v} \in \bar{V}, \bar{u} \neq \bar{v}, N(u) \subsetneq N(v)\}.$$

In such a case we write $\bar{u} \to \bar{v}$. Then $\bar{G} = (\bar{V}, \bar{E})$ is called the ***Hasse diagram*** *or* ***directed rooted tree*** *for graph G.*

Therefore we combine all the vertices who set of neighbors are the same into a class, $\bar{v}$. Note that this is an equivalence relation on V, that is, for $a, b, c \in V$, it is reflexive ($a \sim a$), symmetric ($a \sim b \implies b \sim a$) and transitive ($a \sim b$ and $b \sim c \implies a \sim c$). In Figure 12a, it is clear that $\bar{a} = \{a, c\}, \bar{b} = \{b\}$ and $\bar{d} = \{d\}$. Thus $\bar{V} = \{\bar{a}, \bar{b}, \bar{d}\}$. Also we have that $N(b) \subsetneq N(a)$ and $N(d) \subsetneq N(a)$ which implies that $\bar{b} \to \bar{a}$ and $\bar{d} \to \bar{a}$. The Hasse diagram is shown in Figure 13.

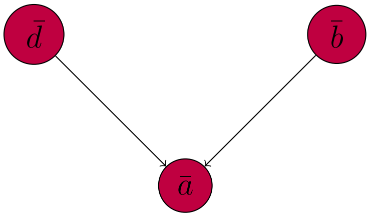


Figure 13: A Hasse diagram for $G = (\{a, b, c, d\}, \{(a, b), (b, c), (a, c), (c, d), (d, a)\})$.

Definition 36. *A* ***tree****, G, is an undirected simple graph that satisfies any of the following:*

1. *G is connected and and has no cycles.*

2. *G has no cycles and a simple graph is formed if any edge is added to G.*

3. *G is connected, but is not connected if any simple edge is removed from G.*

4. *G is connected and K_3 is not a minor of G, i.e. we cannot get K_3 by deleting edges and vertices from G.*

5. *Any 2 vertices in G can be connected by a unique simple path.*

6. *If G has n vertices, then G is connected and has $n - 1$ edges.*

7. *G has no simples cycles and $n - 1$ edges.*

Definition 37. *A* ***directed tree*** *is a tree with directed edges.*

Definition 38. *A* ***rooted tree*** *is a tree where one vertex is designated as the root and all the edges are directed either towards the root or away from the tree.*

Lemma 32. *If G is homogenous, then its Hasse diagram, $\bar{G}$, is a directed rooted tree.*

Thus to find a Hasse ordering for a homogenous graph we need to carry out the followign steps

1. Partition the vertices into equivalence classes.

2. Define a directed rooted tree called a Hasse diagram of G.

Definition 39. *Suppose $G = (V, E)$ with hasse diagram $\bar{G} = (\bar{V}, \bar{E})$. Then $u, v \in \bar{V}$ and $u \to v$ implies that u is an* ***ancestor*** *of v and v is a* ***descentdant*** *of u.*

Definition 40. *An ordering (not necessarily unique) obtained by assigning an ancestor a higher label than any of its descendant is called a* ***Hasse ordering****.*

From Figure 13 we can see that some hasse orderings for Figure 12a would be

$$\begin{pmatrix} a & b & c & d \\ 3 & 1 & 4 & 2 \end{pmatrix} \text{ or } \begin{pmatrix} a & b & c & d \\ 4 & 2 & 3 & 1 \end{pmatrix}.$$

Theorem 41. *Let $G = (\{1, ..., p\}, E)$ be a homogenous graph with a Hasse ordering. Let*

$$\mathbb{P}_G \coloneqq \{\Sigma : \Sigma \in \mathbb{P}^+ \text{ and } (i,j) \notin E \implies \Sigma_{ij} = 0\}$$
$$\mathcal{L}_G \coloneqq \{L : L \text{ is lower triangular and } (i,j) \notin E \implies L_{ij} = 0\}$$

Then

$$\Sigma = LDL^T \in \mathbb{P}_G \iff L \in \mathcal{L}_G \iff L^{-1} \in \mathcal{L}_G$$

Proof. Proof needed. □

The converse is also true.

Theorem 42. *Let $G = (\{1, ..., p\}, E)$ and*

$$\mathbb{P}_G \coloneqq \{\Sigma : \Sigma \in \mathbb{P}^+ \text{ and } (i,j) \notin E \implies \Sigma_{ij} = 0\}$$
$$\mathcal{L}_G \coloneqq \{L : L \text{ is lower triangular and } (i,j) \notin E \implies L_{ij} = 0\}$$

Suppose that

$$\Sigma = LDL^T \in \mathbb{P}_G,$$
$$L \in \mathcal{L}_G \text{ and}$$
$$L^{-1} \in \mathcal{L}_G.$$

Then G is homogenous with a Hasse ordering.

Proof. Proof needed. □

7.6 Closed form of $\hat{\Sigma}_{mle}$

To find $\hat{\Sigma}_{mle}$ we have to solve the following problem:

$$\begin{aligned}\hat{\Sigma}_{mle} &= \arg\min_{\Sigma\in\mathbb{P}_G} l^*(\Sigma)\\ &= \arg\min_{\Sigma\in\mathbb{P}_G} tr(\Sigma^{-1}S) + \log|\Sigma|\end{aligned}$$

Now if G is homogenous, then $\Sigma = LDL^T \implies \Sigma^{-1} = L^{-T}D^{-1}L^{-1}$ where $L \in \mathcal{L}_G$ and

$$D \in \mathcal{D} = \{D : D \text{ is a diangonal matrix such that } \forall i, d_{ii} > 0\}$$

Thus

$$l^*(L, D) = tr(L^{-T}D^{-1}L^{-1}S) + \log|D|$$

Now let $T \coloneqq L^{-1}$. Then

$$\begin{aligned}l^*(T, D) &= tr(T^TD^{-1}TS) + \log|D|\\ &= tr(D^{-1}TST^T) + \log\prod_{i=1}^{p} D_{ii}\\ &= \sum_{i=1}^{p}(D^{-1}TST^T)_{ii} + \sum_{i=1}^{p}\log D_{ii}\\ &= \sum_{i=1}^{p}(\frac{T_{i.}(ST^T)_{.i}}{D_{ii}} + \log D_{ii})\\ &= \sum_{i=1}^{p}(\frac{T_{i.}S(T^T)_{.i}}{D_{ii}} + \log D_{ii})\\ &= \sum_{i=1}^{p}(\frac{T_{i.}ST_{i.}^T}{D_{ii}} + \log D_{ii})\end{aligned}$$

Thus we can minimize with respect to $T_{i.}$ and D_{ii} separately for each i. Note that if $x^i = (t_{ij})_{j<i,(i,j)\in E}$, then

$$\begin{aligned}
T_{i.}ST_{i.}^T &= \sum_{j=1}^{p}\sum_{k=1}^{p} S_{jk}t_{ij}t_{ik} \\
&= \sum_{j=1,(i,j)\in E}^{i} \sum_{k=1,(i,k)\in E}^{i} S_{jk}t_{ij}t_{ik} \qquad \text{as } (i,j)\notin E \implies t_{ij}=0 \\
&= (x^i, 1)\begin{pmatrix} S^{<i} & S_{.i}^{<} \\ (S_{.i}^{<})^T & S_{ii}\end{pmatrix}\begin{pmatrix} x^i \\ 1\end{pmatrix} \\
&= (x^i)^T S^{<i}x^i + (x^i)^T S_{.i}^{<} + (S_{.i}^{<})^T x^i + S_{ii}
\end{aligned}$$

where

$$\begin{aligned}
S_{.i}^{<} &= (S_{k,i})_{k<i,(i,k)\in E} \\
N^{<}(i) &= \{j : j < i, (i,j) \in E\} \\
S^{<i} &= (S_{k,l})_{k,l<i,k,l\in N^{<}(i)}
\end{aligned}$$

Note that to minimize $T_{i.}ST_{i.}^T$ we can simply take the derivative with respect to x^i

$$\begin{aligned}
&\frac{\partial}{\partial x^i}T_{i.}ST_{i.}^T \\
&= \frac{\partial}{\partial x^i}(x^i)^T S^{<i}x^i + (x^i)^T S_{.i}^{<} + (S_{.i}^{<})^T x^i + S_{ii} \\
&= 2S^{<i}x^i + 2S_{.i}^{<} \overset{set}{=} 0 \\
\implies \hat{x}^i &= (S^{<i})^{-1}S_{.i}^{<} \\
\implies \hat{T}_{i.}S\hat{T}_{i.}^T &= (\hat{x}^i)^T S^{<i}\hat{x}^i + (\hat{x}^i)^T S_{.i}^{<} + (S_{.i}^{<})^T \hat{x}^i + S_{ii} \\
&= +(S_{.i}^{<})^T(S^{<i})^{-1}S_{.i}^{<} - (S_{.i}^{<})^T(S^{<i})^{-1}S_{.i}^{<} - (S_{.i}^{<})^T(S^{<i})^{-1}S_{.i}^{<} + S_{ii} \\
&= S_{ii} - (S_{.i}^{<})^T(S^{<i})^{-1}S_{.i}^{<}
\end{aligned}$$

If we define

$$S_i = \begin{pmatrix} S^{<i} & S_{.i}^{<} \\ (S_{.i}^{<})^T & S_{ii}\end{pmatrix},$$

then

$$\begin{aligned} l^*(T,D) &= tr(T^T D^{-1} T S) + \log|D| \\ &= \sum_{i=1}^{p} (\frac{T_{i.} S T_{i.}^T}{D_{ii}} + \log D_{ii}) \\ &= \sum_{i=1}^{p} [\frac{1}{D_{ii}}((x^i, 1) S_i \begin{pmatrix} x^i \\ 1 \end{pmatrix}) + \log D_{ii}]. \end{aligned}$$

Also note that $\frac{\partial}{\partial D_{ii}}(\frac{c}{D_{ii}} + \log D_{ii}) = -\frac{c}{D_{ii}^2} + \frac{1}{D_{ii}} \overset{set}{=} 0 \implies D_{ii} = c$. Thus the value of D_{ii} that minimizes $l^*(T,D)$ is $S_{ii} - (S_{.i}^<)^T (S^{<i})^{-1} S_{.i}^<$.

Lemma 33.

$$\hat{\Sigma} = \Big[\sum_{i=1}^{p} [S_i^{-1}]^0 - \sum_{i=1}^{p} [(S^{<i})^{-1}]^0 \Big]^{-1}$$

Proof. We do the proof for $p = 3$. Let $S = (s_{ij})$. Then we have that

$$\begin{aligned} S_1 &= s_{11} \\ S_2 &= \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \\ S^{<2} &= s_{11} = S_1 \\ S_3 &= S \\ S^{<3} &= \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} = S_2 \end{aligned}$$

Thus

$$\begin{aligned} \hat{\Sigma} &= \Big[\sum_{i=1}^{p} [S_i^{-1}]^0 - \sum_{i=1}^{p} [(S^{<i})^{-1}]^0 \Big]^{-1} \\ &= \Big[[S_1^{-1}]^0 - [(S^{<2})^{-1}]^0 + [S_2^{-1}]^0 - [(S^{<3})^{-1}]^0 + [S_3^{-1}]^0 \Big]^{-1} \\ &= \Big[[S_1^{-1}]^0 - [S_1^{-1}]^0 + [S_2^{-1}]^0 - [S_2^{-1}]^0 + [S_3^{-1}] \Big]^{-1} \\ &= S_3 = S. \end{aligned}$$

□

7.7 Direct sample from $\mathcal{GIW}(U, \delta)$ for homogenous graphs

Recall that the pdf of the $\mathcal{GIW}(U, \delta)$ is proportional to

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

Assume G is homogenous. Then there is a Hasse ordering for the vertices of G and suppose

$$\begin{aligned}
&\Sigma = LDL^T \\
&T = L^{-1}, T \in \mathcal{L}_{\mathcal{G}} \\
\implies &\Sigma \mapsto (L, D) \text{ is a bijection} \\
&(L, D) \mapsto (T, D) \text{ is a bijection}
\end{aligned}$$

where the determinant of $\Sigma \mapsto (L, D)$ is $\prod_{j=1}^{p} D_{jj}^{n_j}, n_j = |\{i : i > j, (i, j) \in E\}|$ and the determinant of $(L, D) \mapsto (T, D)$ is 1. Therefore,

$$\begin{aligned}
\pi_{U,\delta}(T, D) &\propto e^{-\frac{1}{2}tr(T^TD^{-1}TU)-\frac{\delta}{2}\log|D|}\prod_{j=1}^{p} D_{jj}^{n_j} \\
&= e^{-\sum_{i=1}^{p}\frac{(TUT^T)_{ii}}{2D_{ii}}-sum_{i=1}^{p}(\frac{\delta-2n_i}{2})\log D_{ii}}
\end{aligned}$$

7.8 Graphical Lasso

Let

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

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,

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

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

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

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.

7.9 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

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

where

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} \Omega_{1,1} & \Omega_{1,-1} \\ \Omega_{-1,1} & \Omega_{-1,-1} \end{pmatrix}$$

As $\Sigma_{-1,-1}$ is invertible we have by using Schur Complements that

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1} & -(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1} \\ -\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1} & \Sigma_{-1,-1} + \Sigma_{-1,-1}^{-1}\Sigma_{-1,1}(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1}\Sigma_{-1,1}\Sigma_{-1,-1}^{-1} \end{pmatrix}$$

Hence,

$$\begin{aligned} \Omega_{11} &= (\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1} \\ \Omega_{1,-1} &= -(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1} \\ \implies -(\Omega_{1,1})^{-1}\Omega_{1,-1} &= (\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}) \\ &\quad \times (-(\Sigma_{1,1} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1})^{-1}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1}) \\ &= -\Sigma_{1,-1}\Sigma_{-1,-1}^{-1} \end{aligned}$$

This gives us the following lemma:

Lemma 34. $\Sigma_{i,-i}\Sigma_{-i,-i}^{-1} = -(\Omega_{ii})^{-1}\Omega_{i,-i}$.

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

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

and the partial correlation coefficient

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

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:

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

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.

7.10 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

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

Algorithm 1 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 $\{\hat{\Omega}_{ii}^{(r-1)}\}_{i=1}^{p}$)
    for $i = 1, ..., p$ do

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

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

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

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

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 35. *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.

Algorithm 2 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 $i = 1, ...p - 1$ do
        for $j = 1, ...p - 1$ do
            $$\hat{\Omega}_{ij}^{current} = (T_{ij}(\Omega^{current}))_{ij}$$
        end for
    end for                             ▷ updates to partial variances $\Omega_{ii}$
    for $i = 1, ...p - 1$ 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 $(\hat{\Omega^{(r)}})$
```