Assignment3

Yuki Joyama (yj2803)

Problem 1

$X \sim N(\mu, \Sigma)$ where $X = (X_1, ...X_p)$
The conditional distribution $X_i, X_j | X_s = x_s$ is still a Gaussian distribution with mean and covariance given by: $X_i, X_j | X_s = x_s \sim N(\mu_{ij|S}, \Sigma_{ij|S})$

$$\left(\Sigma_{\{ijS\},\{ijS\}}\right)^{-1} = \begin{pmatrix} K_{\{i,j\},\{i,j\}} & K_{\{i,j\},S} \\ K_{S,\{i,j\}} & K_S \end{pmatrix}$$

When $X_i \perp\!\!\!\perp X_j | X_s$ is true, $K_{ij|S} = 0$, which makes $\left(\Sigma_{\{ijS\},\{ijS\}}\right)^{-1}$ diagonal. Therefore, $X_i \perp\!\!\!\perp X_j | X_s \Rightarrow (\Sigma_{\{ijS\},\{ijS\}})^{-1}_{ij} = 0$ holds.
When $(\Sigma_{\{ijS\},\{ijS\}})^{-1}_{ij} = 0$ is true, $K_{ij|S} = (\Sigma_{ij|S})^{-1} = 0$. $(\Sigma_{ij|S})^{-1}$ is the precision matrix for the conditional distribution $X_i, X_j | X_s = x_s$, so X_i and X_j are independent given X_s.
Therefore, $X_i \perp\!\!\!\perp X_j | X_s \Leftrightarrow (\Sigma_{\{ijS\},\{ijS\}})^{-1}_{ij} = 0$.

Problem 2

```r
# simulate data from a given MRF independence model
set.seed(123)
K <- cbind(c(10,7,7,0),c(7,20,0,7),c(7,0,30,7),c(0,7,7,40))
data <- as.data.frame(mvrnorm(n=10000,mu=c(0,0,0,0),Sigma=solve(K)))
colnames(data) <- c("X1","X2","X3","X4")
K
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   10    7    7    0
## [2,]    7   20    0    7
## [3,]    7    0   30    7
## [4,]    0    7    7   40
```

In the precision matrix, $K_{ij} = 0$ implies that variable X_i and X_j are conditionally independent given all other variables. Given the precision matrix K, $K_{14} = K_{41} = 0$ and $K_{23} = K_{32} = 0$. Therefore the following conditional independencies are represented by K:
$X_1 \perp\!\!\!\perp X_4 | X \backslash \{X_1, X_4\}$
$X_2 \perp\!\!\!\perp X_3 | X \backslash \{X_2, X_3\}$ The corresponding graph is an undirected graph that has no edges between X_1 and

X_4, and X_2 and X_3. All other pairs of variables are connected by edges.
Now, I will verify the conditional independence constraints by using linear regression.
$X_1 \perp\!\!\!\perp X_4 | X \setminus \{X_1, X_4\}$:

```r
# conditional independence of X1 and X4 given X2, X3
m14 = lm(X1 ~ X2 + X3 + X4, data = data)
summary(m14)
```

```
##
## Call:
## lm(formula = X1 ~ X2 + X3 + X4, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36729 -0.21127  0.00304  0.21389  1.20994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001934   0.003141   0.616    0.538
## X2          -0.682729   0.012203 -55.950   <2e-16 ***
## X3          -0.695282   0.015540 -44.741   <2e-16 ***
## X4           0.007927   0.020037   0.396    0.692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3141 on 9996 degrees of freedom
## Multiple R-squared:  0.4564, Adjusted R-squared:  0.4563
## F-statistic:  2798 on 3 and 9996 DF,  p-value: < 2.2e-16
```

In this linear model, the coefficient of $X4$ turned out to be non-significant with p-value <0.05.

$X_2 \perp\!\!\!\perp X_3 | X \setminus \{X_2, X_3\}$:

```r
# conditional independence of X2 and X3 given X1, X4
m23 = lm(X2 ~ X1 + X3 + X4, data = data)
summary(m23)
```

```
##
## Call:
## lm(formula = X2 ~ X1 + X3 + X4, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90282 -0.15318  0.00188  0.15342  0.85952
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001141   0.002247   0.508    0.612
## X1          -0.349303   0.006243 -55.950   <2e-16 ***
## X3           0.012316   0.012177   1.011    0.312
## X4          -0.352810   0.013891 -25.398   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2246 on 9996 degrees of freedom
## Multiple R-squared:  0.3841, Adjusted R-squared:  0.3839
## F-statistic:  2078 on 3 and 9996 DF,  p-value: < 2.2e-16
```

In this linear model, the coefficient of $X3$ turned out to be non-significant with p-value <0.05.

Therefore, the conditional independencies are verified.

The list of edges are $X_1 - X_2$, $X_1 - X_3$, $X_2 - X_4$, $X_3 - X_4$.

```r
# fit the model (estimate the precision matrix subject to the graph constraints)
library(gRim)
glist <- list(
  c("X1", "X2"),
  c("X1", "X3"),
  c("X2", "X4"),
  c("X3", "X4")
)
ddd <- cov.wt(data, method="ML")
fit <- ggmfit(ddd$cov, ddd$n.obs, glist) # Estimate parameters using IPF
fit$K # estimated precision matrix
```

```
##           X1        X2        X3        X4
## X1 10.182411  6.988142  7.140856  0.000000
## X2  6.988142 19.832337  0.000000  7.076402
## X3  7.140856  0.000000 29.394792  6.852069
## X4  0.000000  7.076402  6.852069 40.745105
```

It appears that the model fitting worked because we can see that the estimated precision matrix has $K_{14} = K_{41} = 0$ and $K_{23} = K_{32} = 0$, and everything else non-zero, indicating that the above conditional independencies hold.

Problem 3

```r
# Gaussian Bayesian Network model
# covariance matrix
```

```r
set.seed(123)
Sig <- cbind(c(3,-1.4,0,0),c(-1.4,3,1.4,1.4),c(0,1.4,3,0),c(0,1.4,0,3))
data <- as.data.frame(mvrnorm(n=10000,mu=c(0,0,0,0),Sigma=Sig))
colnames(data) <- c("X1","X2","X3","X4")
```

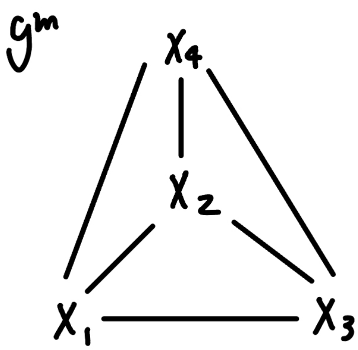
DAG $\mathcal{G}$: $X_1 \rightarrow X_2 \leftarrow X_3$ and $X_4 \rightarrow X_2$

(a) Given the model, correlation constraints are:
$X_1 \perp\!\!\!\perp X_3 | X_2$
$X_1 \perp\!\!\!\perp X_4 | X_2$
$X_3 \perp\!\!\!\perp X_4 | X_2$

```r
# estimate the correlation
cor(data)
```

```
##             X1         X2          X3           X4
## X1  1.00000000 -0.4661188 0.012187930 -0.011504563
## X2 -0.46611880  1.0000000 0.463034923  0.473314198
## X3  0.01218793  0.4630349 1.000000000  0.006392376
## X4 -0.01150456  0.4733142 0.006392376  1.000000000
```

We can see that the correlation between X_1 and X_3, X_1 and X_4, X_3 and X_4 are very close to zero.



(b)
The precision matrix K for $\mathcal{G}^m$ is 4 by 4 matrix with non-zero off-diagonal elements because all the pairs in the moralized graph are connected by edges. Additionally, K indicates that

(c)

Problem 4

```r
library(dagitty)

# simulate 10000 observations from the following graph
g <- dagitty( "dag{ x <- u1; u1 -> m <- u2 ; u2 -> y }" )
```

Problem 5