P8124 Assignment 3

Christine Lucille Kuryla (clk2162)

2023-10-23

Problem 2

Simulate data from given MRF independence model

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```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)) )
```

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

```r
data <- as.data.frame(mvrnorm(n=10000,mu=c(0,0,0,0),Sigma=solve(K)))
colnames(data) <- c("X1","X2","X3","X4")

# (Note: in R, the inverse of a matrix M is obtained by solve(M).)
```

Conditional Independencies

What are the conditional independencies that are representing in this precision matrix? Conditional independencies correspond to the zeros in the precision matrix of the elements given everything else. Hence, for K, the conditional independencies are:

$X_1 \perp X_4 | X_2, X_3$

and

$X_2 \perp X_3 | X_1, X_4$

Corresponding Graph (INSERT PIC?!?!?!?!)

What is the corresponding graph? The corresponding MRF has vertices X_1, X_2, X_3, X_4 and edges:

- $X_1 - X_2$.

- $X_2 - X_4$.
- $X_4 - X_3$.
- $X_3 - X_1$.

Verify with linear regression

Verify the conditional independence constraints by using linear regression.

```r
# X1 independent of X4 given X2, X3
summary(glm(data = data, formula = X1 ~ X4 + X2 + X3))
```

```
##
## Call:
## glm(formula = X1 ~ X4 + X2 + X3, data = data)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001934   0.003141   0.616    0.538
## X4           0.007927   0.020037   0.396    0.692
## X2          -0.682729   0.012203 -55.950   <2e-16 ***
## X3          -0.695282   0.015540 -44.741   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.09863182)
##
##     Null deviance: 1813.81  on 9999  degrees of freedom
## Residual deviance:  985.92  on 9996  degrees of freedom
## AIC: 5221.2
##
## Number of Fisher Scoring iterations: 2
```

```r
# X2 independent of X3 given X1, X4
summary(glm(data = data, formula = X2 ~ X3 + X1 + X4))
```

```
##
## Call:
## glm(formula = X2 ~ X3 + X1 + X4, data = data)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001141   0.002247   0.508    0.612
## X3           0.012316   0.012177   1.011    0.312
## X1          -0.349303   0.006243 -55.950   <2e-16 ***
## X4          -0.352810   0.013891 -25.398   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.05046277)
##
##     Null deviance: 818.95  on 9999  degrees of freedom
## Residual deviance: 504.43  on 9996  degrees of freedom
## AIC: -1480.4
##
## Number of Fisher Scoring iterations: 2
```

As demonstrated in the first linear regression, $X_1 \perp X_4 | X_2, X_3$ because we can see that when regressing X_1 on X_4, X_2, X_3 gives a large p-value for X_4 because they are conditionally independent since X_2 and X_3 are given (note that their small p values demonstrate that they are dependent). The same logic applies to the second regression for $X_2 \perp X_3 | X_1, X_4$ by regressing X_2 on the rest of the variables and observing a large p-value for X_3, showing independence, because X_1 and X_4 are conditioned on by putting them in the regression.

Explanation

The zeroes in the precision matrix K correspond to the conditional independencies described above. The MRF is the UG with X_1, X_2, X_3, X_4 that has the edge between X_1 and X_4 removed because of the conditional independence $X_1 \perp X_4 | X_2, X_3$ and the edge between X_2 and X_3 removed because of the conditional independence $X_2 \perp X_3 | X_1, X_4$ that were demonstrated by the zeroes in the precision matrix. The linear regression demonstrates that the conditional independencies are true because when one variable is regressed on the rest, the p-value for the variable that it is conditionally independent of is large (because they are independent), and the p-values of the variables in the conditioning set are small (because they are dependent).

Estimate precision matrix subject to graph constraints

```
# Use the gRim package to fit the model, i.e., estimate the precision matrix subject to the graph const

library(gRim)
```

```
## Loading required package: gRbase
```

```
glist <- list( c("X1","X2"), c("X2","X4"), c("X4","X3"), c("X3","X1") )
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
```

```
# Did it work? How do you know?

# Precision matrix (K)
kable(K)
```

10	7	7	0
7	20	0	7
7	0	30	7
0	7	7	40

```
# Estimated precision matrix
kable(fit$K)
```

	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

Yes, it worked. We know this because the estimated precision matrix has the expected zeroes that correspond to the conditional independencies, and in general, the values are quite close to K so a good estimation of the actual precision matrix.

Problem 3

Consider the Gaussian Bayesian Network model with the following covariance matrix: and the DAG G with edges X1 → X2 ← X3 and X4 → X2.

a) Correlation constraints and correlation matrix

- a) What correlation constraints does this model represent? Estimate the correlation matrix. * This model represents three marginal independencies (six correlations shown by the 0s).
- $X_4 \perp X_3$ (X_4 is marginally independent of X_3, so the correlation between X_4 and $X_3 = 0$). Correlations are symmetric, so $corr(X_3, X_4) = corr(X_4, X_3) = 0$.
- $X_1 \perp X_3$ (X_1 is marginally independent of X_3, so the correlation between X_1 and $X_3 = 0$). Correlations are symmetric, so $corr(X_3, X_1) = corr(X_1, X_3) = 0$.
- $X_1 \perp X_4$ (X_1 is marginally independent of X_4, so the correlation between X_1 and $X_4 = 0$). Correlations are symmetric, so $corr(X_4, X_1) = corr(X_1, X_4) = 0$.

```
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)) )
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  3.0 -1.4  0.0  0.0
## [2,] -1.4  3.0  1.4  1.4
## [3,]  0.0  1.4  3.0  0.0
## [4,]  0.0  1.4  0.0  3.0
```

```
data <- as.data.frame(mvrnorm(n=10000,mu=c(0,0,0,0),Sigma=Sig))
colnames(data) <- c("X1","X2","X3","X4")

# Estimate correlation matrix
sigma_est <- cor(data)

kable(sigma_est)
```

	X1	X2	X3	X4
X1	1.0000000	-0.4661188	0.0121879	-0.0115046
X2	-0.4661188	1.0000000	0.4630349	0.4733142
X3	0.0121879	0.4630349	1.0000000	0.0063924
X4	-0.0115046	0.4733142	0.0063924	1.0000000

b) The moralized graph

- b) Consider also the moralized graph Gm and what the corresponding precision matrix K would look like. What are the partial correlation constraints represented in K? How does this make sense with respect to sigma above? *
- The moralized Graph Gm would be the complete graph formed from the skeleton of G. It is the graph formed by making the edges in G undirected and adding edges $X_1 - X_4$, $X_4 - X_3$, and $X_3 - X_1$ because X_2 is an unshielded collider so it's parents are married during the moralization process.

- There are no partial correlation constraints represented in K because there are no missing edges in Gm. XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- This makes sense wrt the correlation matrix Sigma above because there are marginal independencies but no conditional independencies. XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

c) Estimate K, take inverse, and compare to true Sigma

- c) Following steps similar to the previous problem, estimate the corresponding precision matrix K from this data (using ggmfit). Take the inverse and compare to the true covariance matrix. *

```
glist <- list( c("X1","X2"), c("X2","X3"), c("X4","X2")  )
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 0.4270365  0.2001141  0.0000000  0.0000000
## X2 0.2001141  0.6213527 -0.1989045 -0.2021127
## X3 0.0000000 -0.1989045  0.4290929  0.0000000
## X4 0.0000000 -0.2021127  0.0000000  0.4188167
```

```
solve(fit$K) # inverse of K (covariance matrix)
```

```
##            X1        X2         X3         X4
## X1  2.9917221 -1.387081 -0.6429763 -0.6693778
## X2 -1.3870808  2.959982  1.3720889  1.4284287
## X3 -0.6429763  1.372089  2.9665248  0.6621430
## X4 -0.6693778  1.428429  0.6621430  3.0770111
```

Problem 4

```
library(dagitty)
```

```
##
## Attaching package: 'dagitty'

## The following object is masked from 'package:gRim':
##
##     ciTest

## The following objects are masked from 'package:gRbase':
##
##     ancestors, children, edges, moralize, parents
```

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

sim_sem <- simulateSEM(g,
            b.lower = 0.4,
            b.upper = 0.7,
            N = 10000)

# Here U1,U2 represent unmeasured variables.

# Estimate the effect of X on Y adjusting for M in a linear regression, obtaining a 95% confidence inte
```

```r
lm_m = lm(data = sim_sem, formula = y ~ x + m )
summary(lm_m)
```

```
##
## Call:
## lm(formula = y ~ x + m, data = sim_sem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7658 -0.6043  0.0072  0.6044  3.4050
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.008944   0.008972  -0.997    0.319
## x           -0.183864   0.009566 -19.221   <2e-16 ***
## m            0.494160   0.009698  50.956   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8972 on 9997 degrees of freedom
## Multiple R-squared:  0.2062, Adjusted R-squared:  0.2061
## F-statistic:  1299 on 2 and 9997 DF,  p-value: < 2.2e-16
```

```r
library(broom)
tidy_ci_m <- tidy(lm_m, conf.int=TRUE)
tidy_ci_m
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept) -0.00894   0.00897    -0.997 3.19e- 1  -0.0265   0.00864
## 2 x           -0.184     0.00957   -19.2   7.06e-81  -0.203   -0.165
## 3 m            0.494     0.00970    51.0   0          0.475    0.513
```

```r
# Then estimate the same effect (and confidence interval) using the correct sufficient adjustment set t

# Sufficient adjustment set
adjustmentSets(g, "y", "x", type = "minimal")
```

```
##  {}
```

```r
# This results in an empty set.



##### HOW?



# What conclusion should be drawn from this example?
```

The estimate of the effect of X on Y adjusting for M in a linear regression is -0.1838637, with a confidence interval (-0.2026147, -0.1651126).

The sufficient adjus

The estimate

Problem 5

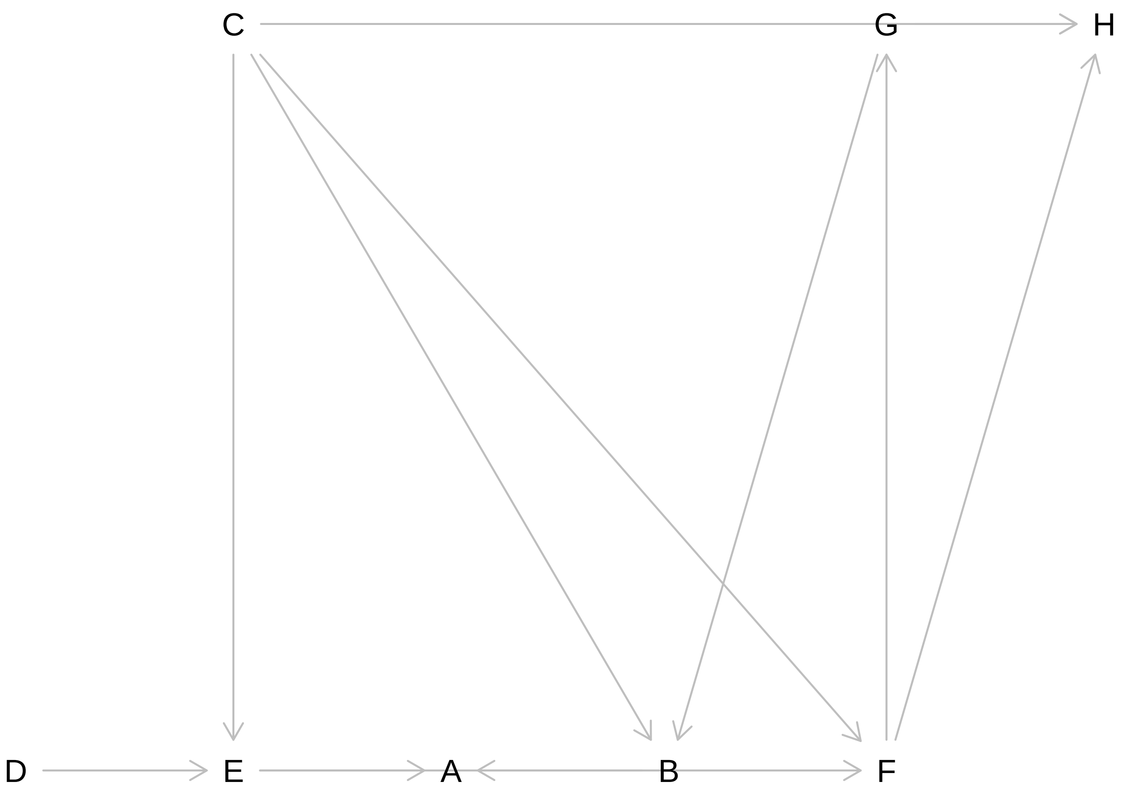
```r
# Construct the DAG in Figure 1 as a daggity object.

dag_q5 <- dagitty('dag {
    D [pos="0,1"]
    E [pos="1,1"]
    A [pos="2,1"]
    B [pos="3,1"]
    F [pos="4,1"]
    C [pos="1,0"]
    G [pos="4,0"]
    H [pos="5,0"]

    D -> E -> A <- B <- G -> H
    E -> F -> H
    E <- C -> H
    C -> B
    C -> F -> G

}')

plot(dag_q5)
```




```r
# Simulate 10000 observations from this graph using simulateSEM() as you did on the first homework.

sim_sem <- simulateSEM(dag_q5,
            b.lower = -0.7,
            b.upper = 0.7,
```

```r
            N = 10000)

# Estimate the effect of E on F and the effect of B on A using backdoor adjustment and linear regressio
# ????????????????????????

# Effect of E on F
adjustmentSets(dag_q5, "E", "F", type = "minimal")
```

```
## { C }
```

```r
# Result: { C }

# Effect of B on A
adjustmentSets(dag_q5, "B", "A", type = "minimal")
```

```
## { E }
## { C, F }
## { C, G }
```

```r
# Result: { E }, { C, F }, { C, G }

# If there is more than one sufficient adjustment set, try each of the ones identified by dagitty and c

# Are the point estimates similar? Do the estimates have similar variance (or confidence interval lengt
```