

MATH 598/782 – PROJECT 1 SOLUTIONS

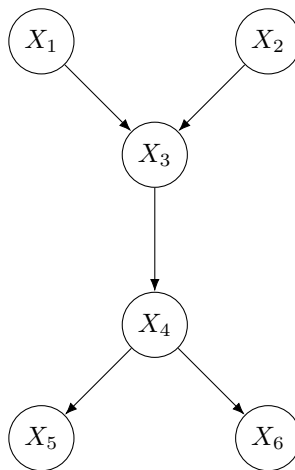
For each of the following causal DAGs that define a joint distribution for a collection of random variables,

- (a) write out the implied factorization of the joint distribution;
- (b) simulate data from the joint distribution (you may make any suitable distributional assumptions);
- (c) using the simulated data, demonstrate one example of unconditional d-separation, and one example of conditional d-separation; if no such examples exist, state this. It may be sufficient to demonstrate uncorrelatedness rather than independence.

Solution: We will stick to Gaussian simulation throughout.

1. DAG 1:

5 MARKS



(a) Factorization:

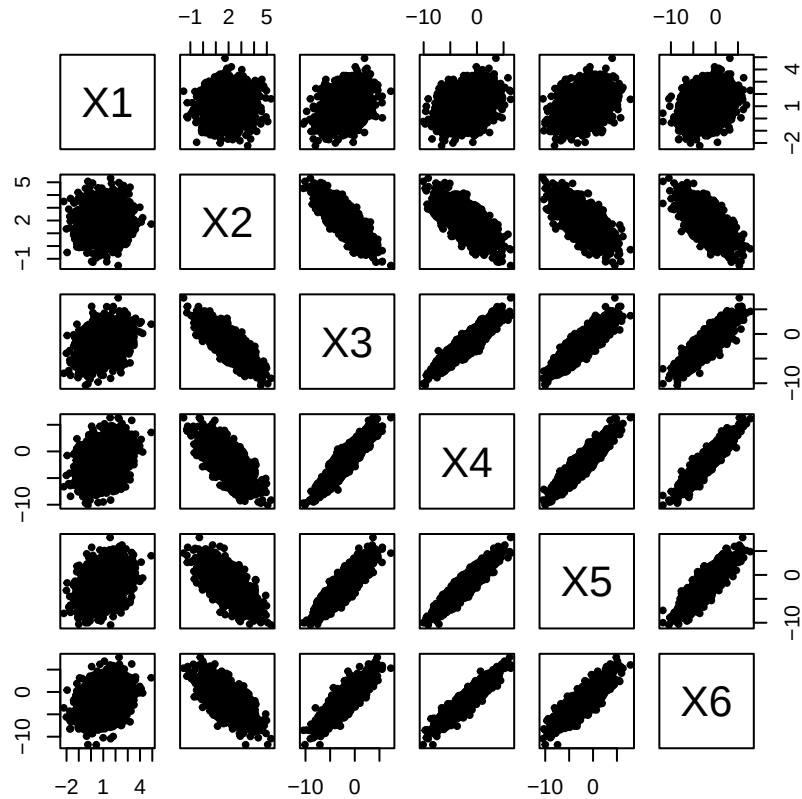
$$f_{X_1}(x_1)f_{X_2}(x_2)f_{X_3|X_1,X_2}(x_3|x_1,x_2)f_{X_4|X_3}(x_4|x_3)f_{X_5|X_4}(x_5|x_4)f_{X_6|X_4}(x_6|x_4)$$

(b) Simulation

```
set.seed(238)
n<-2000
X1<-rnorm(n,1,1)
X2<-rnorm(n,2,1)
X3<-rnorm(n,1+X1-2*X2,1)
X4<-rnorm(n,X3,1)
X5<-rnorm(n,X4,1)
X6<-rnorm(n,X4,1)
Xm<-cbind(X1,X2,X3,X4,X5,X6)
round(cor(Xm),4)

+      X1      X2      X3      X4      X5      X6
+ X1 1.0000  0.0316  0.3838  0.3620  0.3450  0.3360
+ X2 0.0316  1.0000 -0.8128 -0.7465 -0.7004 -0.6949
+ X3 0.3838 -0.8128  1.0000  0.9231  0.8655  0.8600
+ X4 0.3620 -0.7465  0.9231  1.0000  0.9340  0.9304
+ X5 0.3450 -0.7004  0.8655  0.9340  1.0000  0.8688
+ X6 0.3360 -0.6949  0.8600  0.9304  0.8688  1.0000

par(pty='s',mar=c(3,2,1,0))
pairs(Xm,pch=19,cex=0.6)
```



(c) d-separation:

- (i) Unconditional d-separation: only one pair is unconditionally d-separated, X_1 and X_2 ; we can observe that they are uncorrelated with correlation 0.0316455, which we can confirm empirically using the regression of one on the other

```
summary(lm(X2~X1))$coef
+               Estimate Std. Error  t value Pr(>|t|)
+ (Intercept)  1.93827365  0.03141716 61.694753 0.0000000
+ X1           0.03115744  0.02201577  1.415233 0.1571561
```

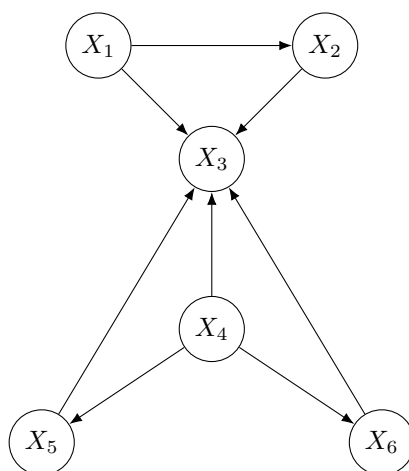
- (ii) Conditional d-separation: There are many examples, but for example X_1 is conditionally d-separated from X_5 by both X_3 and X_4 ; we can assess that using regression

```
summary(lm(X5~X1+X3))$coef
+               Estimate Std. Error  t value  Pr(>|t|)
+ (Intercept) -0.10124604  0.05943340 -1.703521 0.08862625
+ X1           0.04103552  0.03316040  1.237486 0.21605230
+ X3           0.98704227  0.01392746 70.870230 0.00000000

summary(lm(X5~X1+X4))$coef
+               Estimate Std. Error  t value  Pr(>|t|)
+ (Intercept) -0.05803340  0.041050413 -1.4137104 0.1576028
+ X1           0.02148514  0.023443110  0.9164798 0.3595260
+ X4           0.99449470  0.009160445 108.5640116 0.0000000
```

2. DAG 2:

5 MARKS



(a) Factorization: X_1 and X_4 are founders

$$f_{X_1}(x_1)f_{X_4}(x_4)f_{X_2|X_1}(x_2|x_1)f_{X_5|X_4}(x_5|x_4)f_{X_6|X_4}(x_6|x_4)f_{X_3|X_1,X_2,X_4,X_5,X_6}(x_3|x_1,x_2,x_4,x_5,x_6)$$

(b) Simulation:

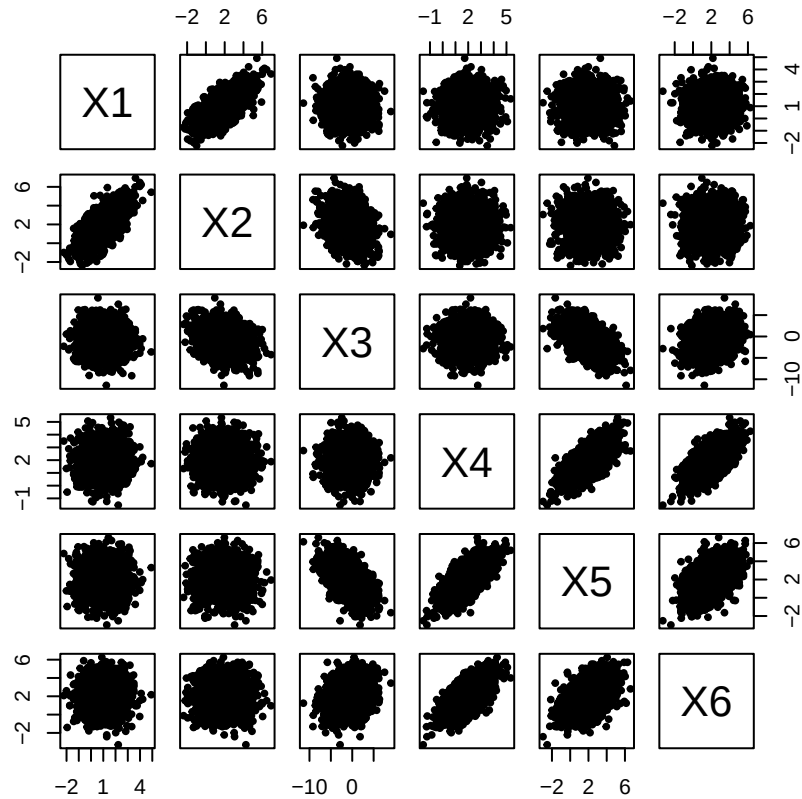
```

set.seed(238)
n<-2000
X1<-rnorm(n,1,1)
X4<-rnorm(n,2,1)
X2<-rnorm(n,1+X1,1)
X5<-rnorm(n,X4,1)
X6<-rnorm(n,X4,1)
X3<-rnorm(n,X1-X2+X4-2*X5+X6,1)
Xm<-cbind(X1,X2,X3,X4,X5,X6)
round(cor(Xm),4)

+           X1           X2           X3           X4           X5           X6
+ X1 1.0000  0.7034  0.0060  0.0316  0.0312  0.0343
+ X2 0.7034  1.0000 -0.2587  0.0245  0.0192  0.0178
+ X3 0.0060 -0.2587  1.0000 -0.0239 -0.5452  0.2625
+ X4 0.0316  0.0245 -0.0239  1.0000  0.7159  0.7085
+ X5 0.0312  0.0192 -0.5452  0.7159  1.0000  0.4947
+ X6 0.0343  0.0178  0.2625  0.7085  0.4947  1.0000

par(pty='s',mar=c(3,2,1,0))
pairs(Xm,pch=19,cex=0.6)

```



(c) d-separation:

- (i) Unconditional d-separation: as X_3 acts as a collider on many paths, X_1 and X_2 are both unconditionally d-separated from each of X_4 , X_5 and X_6 . We can see the zeros in the correlation matrix, and also confirm by regression, for example:

```
summary(lm(X6~X1))$coef
+               Estimate Std. Error  t value      Pr(>|t|)
+ (Intercept)  1.89628395  0.04416217  42.939099  5.744446e-286
+ X1           0.04754186  0.03094692   1.536239  1.246382e-01
summary(lm(X5~X2))$coef
+               Estimate Std. Error  t value      Pr(>|t|)
+ (Intercept)  1.92243040  0.05490069  35.0165057  7.054595e-210
+ X2           0.01953221  0.02272359   0.8595563  3.901368e-01
```

- (ii) Conditional d-separation: There are several examples, but for example X_5 is conditionally d-separated from X_6 by X_4 ; the other path is blocked by the collider X_3 .

```
summary(lm(X5~X6+X4))$coef
+               Estimate Std. Error  t value      Pr(>|t|)
+ (Intercept) -0.05853784  0.04932399  -1.186803  2.354466e-01
+ X6          -0.02565220  0.02254014  -1.138068  2.552287e-01
+ X4           1.05055320  0.03168684  33.154244  2.043479e-192
```

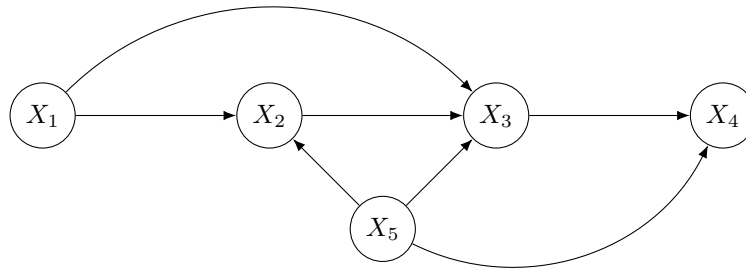
If we add the collider into the model, the dependence of X_5 on X_6 becomes non negligible.

```
summary(lm(X5~X6+X3+X4))$coef
+               Estimate Std. Error  t value      Pr(>|t|)
+ (Intercept) -0.3212728  0.029013753  -11.07312  1.065529e-27
+ X6           0.3277657  0.014287358   22.94096  1.446951e-103
+ X3          -0.3315231  0.005308992  -62.44558  0.000000e+00
```

```
+ X4          0.6773486 0.019386281 34.93959 4.132891e-209
```

3. DAG 3:

5 MARKS



(a) Factorization: X_1 and X_5 are founders

$$f_{X_1}(x_1)f_{X_5}(x_5)f_{X_2|X_1,X_5}(x_2|x_1,x_5)f_{X_3|X_1,X_2,X_5}(x_3|x_1,x_2,x_5)f_{X_4|X_3,X_5}(x_4|x_3,x_5)$$

(b) Simulation:

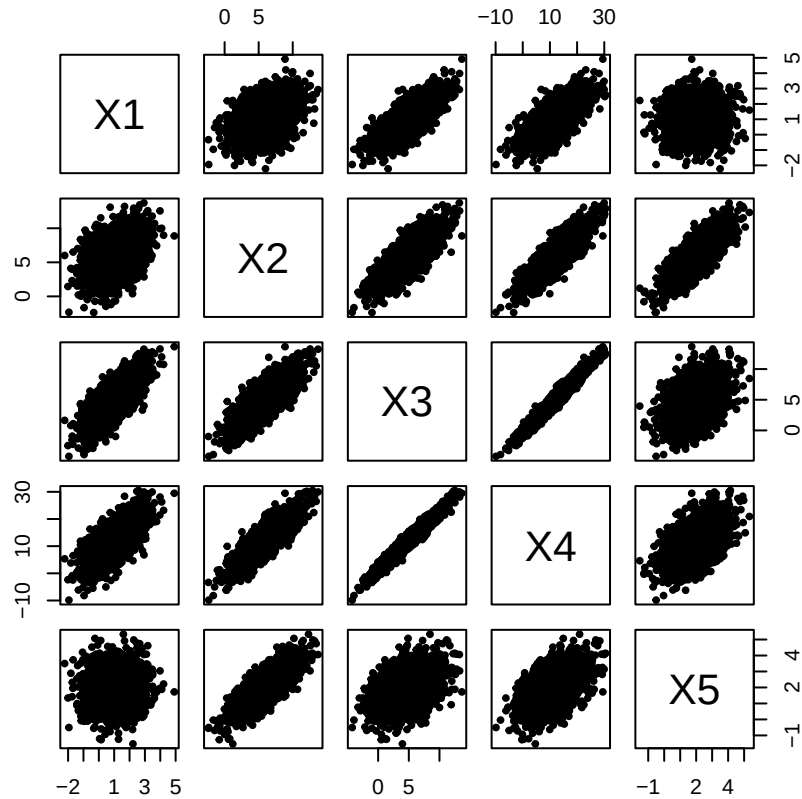
```

set.seed(238)
n<-2000
X1<-rnorm(n,1,1)
X5<-rnorm(n,2,1)
X2<-rnorm(n,1+X1+2*X5,1)
X3<-rnorm(n,X1+X2-X5,1)
X4<-rnorm(n,2*X3+X5,1)
Xm<-cbind(X1,X2,X3,X4,X5)
round(cor(Xm),4)

+          X1      X2      X3      X4      X5
+ X1 1.0000 0.4261 0.7620 0.6988 0.0316
+ X2 0.4261 1.0000 0.7827 0.8473 0.8221
+ X3 0.7620 0.7827 1.0000 0.9739 0.4055
+ X4 0.6988 0.8473 0.9739 1.0000 0.5352
+ X5 0.0316 0.8221 0.4055 0.5352 1.0000

par(pty='s',mar=c(3,2,1,0))
pairs(Xm,pch=19,cex=0.6)

```



(c) d-separation:

(i) Unconditional d-separation: only X_1 and X_5 are unconditionally d-separated.

```
summary(lm(X5~X1))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept)  1.93827365  0.03141716  61.694753  0.0000000
+ X1           0.03115744  0.02201577   1.415233  0.1571561
```

(ii) Conditional d-separation: There are several examples, for instance X_2 is conditionally d-separated from X_4 by (X_3, X_5) .

```
summary(lm(X4~X2+X3+X5))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept) -0.018857790  0.05743852  -0.3283126  7.427098e-01
+ X2           0.008900234  0.03124239   0.2848768  7.757681e-01
+ X3           1.994386322  0.01788405  111.5176150  0.000000e+00
+ X5           0.984037584  0.05264621   18.6915196  5.659393e-72
```

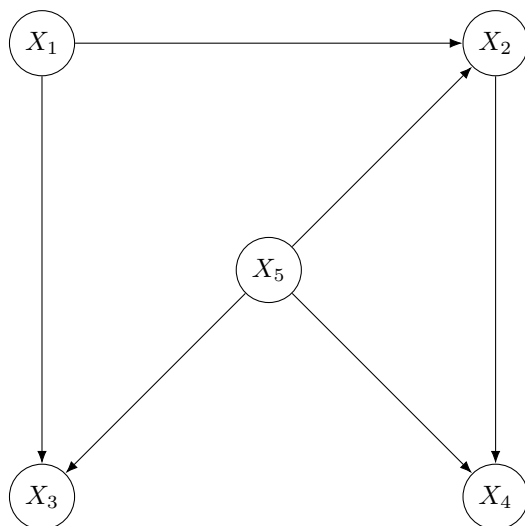
Conditioning on X_3 alone, or X_5 alone, is not sufficient to yield d-separation.

```
summary(lm(X4~X2+X3))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept) -0.03283715  0.06224197  -0.5275725  5.978547e-01
+ X2           0.52698866  0.01562244  33.7327944  8.304642e-198
+ X3           1.76986883  0.01435908  123.2578187  0.000000e+00
```

```
summary(lm(X4~X2+X5))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept) -0.174768  0.15436593  -1.132167  2.577003e-01
+ X2           3.017159  0.04236963  71.210403  0.000000e+00
+ X5          -2.959176  0.10485469 -28.221685  9.491900e-148
```

4. DAG 4:

5 MARKS

(a) Factorization: X_1 and X_5 are founders

$$f_{X_1}(x_1)f_{X_5}(x_5)f_{X_2|X_1,X_5}(x_2|x_1,x_5)f_{X_3|X_1,X_5}(x_3|x_1,x_5)f_{X_4|X_2,X_5}(x_4|x_2,x_5)$$

(b) Simulation:

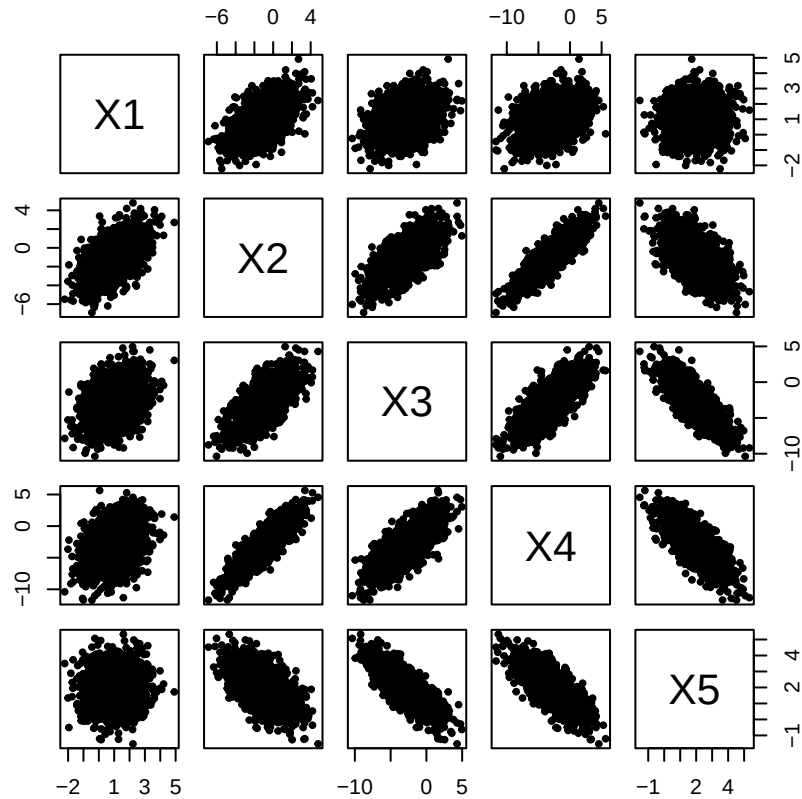
```

set.seed(238)
n<-2000
X1<-rnorm(n,1,1)
X5<-rnorm(n,2,1)
X2<-rnorm(n,X1-X5,1)
X3<-rnorm(n,X1-2*X5,1)
X4<-rnorm(n,X2-X5)
Xm<-cbind(X1,X2,X3,X4,X5)
round(cor(Xm),4)

+           X1          X2          X3          X4          X5
+ X1 1.0000  0.5619  0.3998  0.3609  0.0316
+ X2 0.5619  1.0000  0.6996  0.8696 -0.5653
+ X3 0.3998  0.6996  1.0000  0.7639 -0.8047
+ X4 0.3609  0.8696  0.7639  1.0000 -0.7525
+ X5 0.0316 -0.5653 -0.8047 -0.7525  1.0000

par(pty='s',mar=c(3,2,1,0))
pairs(Xm,pch=19,cex=0.6)

```



(c) d-separation:

(i) Unconditional d-separation: only X_1 and X_5 are unconditionally d-separated.

```
summary(lm(X5~X1))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept)  1.93827365  0.03141716  61.694753  0.0000000
+ X1           0.03115744  0.02201577   1.415233  0.1571561
```

(ii) Conditional d-separation: X_2 and X_3 are conditionally d-separated given (X_1, X_5)

```
summary(lm(X3~X2+X1+X5))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept) -0.069682322  0.05357684  -1.3006052  1.935438e-01
+ X2          -0.008725431  0.02233321  -0.3906932  6.960657e-01
+ X1           1.020559945  0.03099513  32.9264661  2.775457e-190
+ X5          -1.984081440  0.03156867 -62.8496974  0.000000e+00
```

Conditioning on X_1 alone, or X_5 alone, is not sufficient to yield d-separation.

```
summary(lm(X3~X2+X1))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept) -1.98109953  0.07611119 -26.0290172  7.471432e-129
+ X2           0.98109201  0.02732375  35.9061979  2.856443e-218
+ X1           0.02350712  0.04594713   0.5116123  6.089790e-01
```

```
summary(lm(X3~X2+X5))$coef
+               Estimate Std. Error   t value   Pr(>|t|)
+ (Intercept)  0.4301884  0.06381154   6.741546  2.044385e-11
+ X2           0.5083630  0.01972073  25.778102  9.757886e-127
+ X5          -1.4520675  0.03368153 -43.111680  1.326384e-287
```