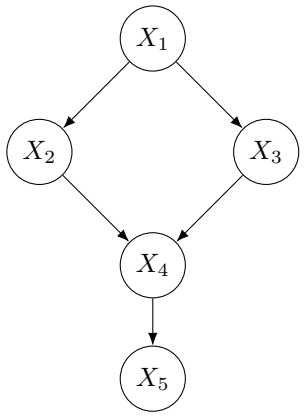
# PROJECT 1: SOLUTIONS

We use the DAG




which implies the factorization

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

with

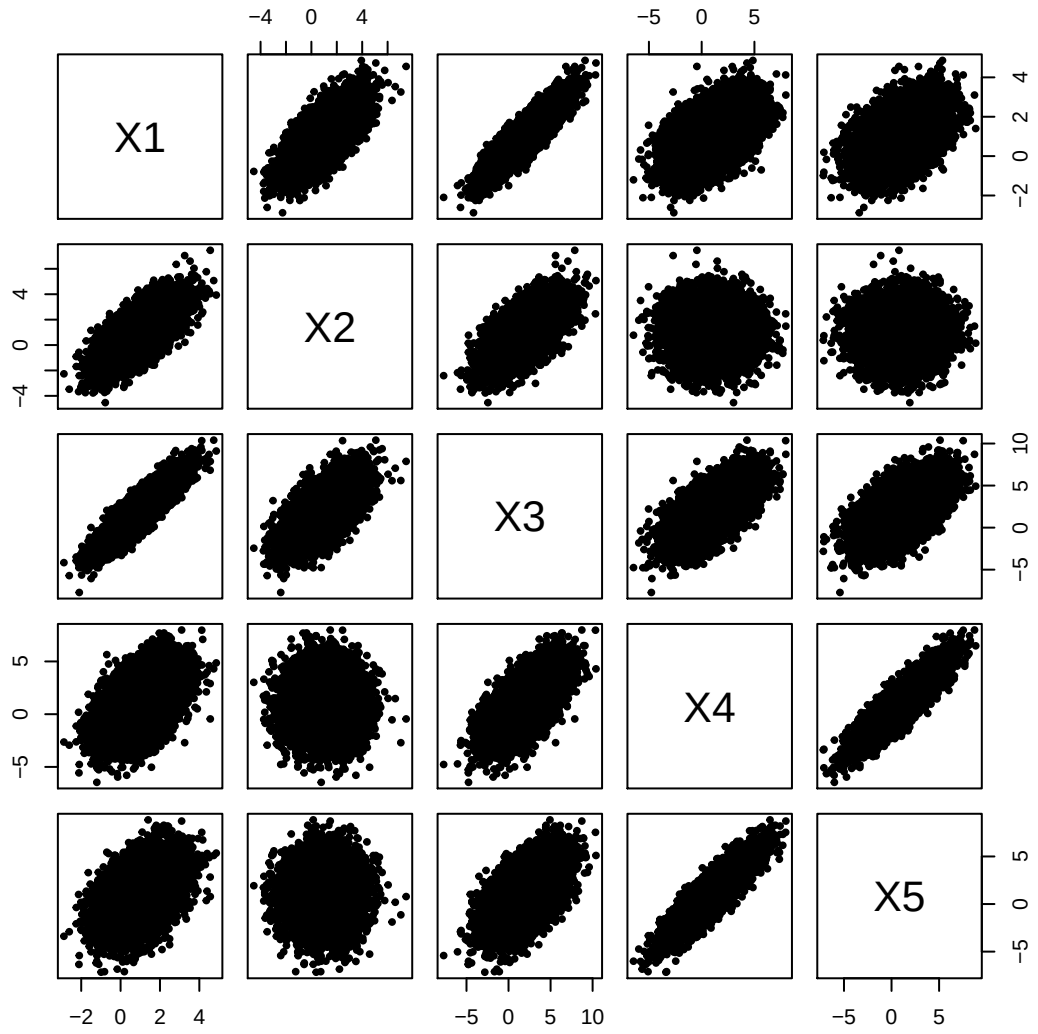
- $X_1 \sim Normal(1,1)$
- $X_2|X_1 = x_1 \sim Normal(x_1,1)$ and $X_3|X_1 = x_1 \sim Normal(2x_1,1)$ independent
- $X_4|X_2 = x_2, X_3 = x_3 \sim Normal(x_3 - x_2, 1)$
- $X_5|X_4 = x_4 \sim Normal(x_4,1)$

yielding a jointly multivariate Normal distribution.

```
#Calculation for large N
library(mvnfast)
set.seed(22087)
n<-10000
X1<-rnorm(n,1,1)
X2<-rnorm(n,X1,1)
X3<-rnorm(n,2*X1,1)
X4<-rnorm(n,X3-X2,1)
X5<-rnorm(n,X4,1)
X<-cbind(X1,X2,X3,X4,X5)
cor(X)                    #Correlation

+           X1           X2        X3           X4          X5
+ X1 1.0000000  0.709102977 0.8937282  0.497721721 0.447904499
+ X2 0.7091030  1.000000000 0.6342668 -0.001229615 0.004090683
+ X3 0.8937282  0.634266801 1.0000000  0.673757466 0.604696876
+ X4 0.4977217 -0.001229615 0.6737575  1.000000000 0.894530465
+ X5 0.4479045  0.004090683 0.6046969  0.894530465 1.000000000

par(mar=c(2,2,2,2))
pairs(X,pch=19,cex=0.6)
```




```
Xa<-X
```

(a) **Blocking and unblocking via conditioning:** By construction, $X_2$ and $X_3$ are conditionally independent given $X_1$, but are marginally dependent . We can demonstrate this by regression, or by indpecting the correlation or partial correlation between the variables: first, for dependence:

```
cor(X2,X3)                  #Correlation not zero

+ [1] 0.6342668

coef(summary(lm(X3~X2)))    #Regression coefficient not zero

+             Estimate Std. Error  t value Pr(>|t|)
+ (Intercept) 1.028771 0.02114754 48.64728        0
+ X2          1.001631 0.01221021 82.03230        0

coef(summary(lm(X2~X3)))    #Regression coefficient not zero

+              Estimate  Std. Error  t value     Pr(>|t|)
+ (Intercept) 0.1835860 0.014778514 12.42249 3.584674e-35
+ X3          0.4016392 0.004896111 82.03230 0.000000e+00
```

Secondly, for conditional independence, we can inspect the partial correlation of $(X_2, X_3)$ given $X_1$ by looking at the $(2,3)$ element of the inverse correlation matrix formed by considering the three variables.

```
solve(cor(cbind(X1,X2,X3))) #almost zero in the (2,3) element

+           X1           X2           X3
+ X1  5.973724 -1.421615807 -4.437201678
+ X2 -1.421616  2.011377895 -0.005212151
+ X3 -4.437202 -0.005212151  4.968957962

coef(summary(lm(X3~X2+X1))) #conditioning on X1 makes the coefficient for X2 almost zero.

+                Estimate Std. Error     t value  Pr(>|t|)
+ (Intercept) 0.004667418 0.01427759   0.3269051 0.7437465
+ X2          0.001656485 0.01004881   0.1648439 0.8690702
+ X1          2.001939110 0.01426546 140.3347118 0.0000000
```

(b) **d-separation and conditional independence:** From the DAG, we can see that

$$X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3)$$

so that $X_1$ and $X_4$ are d-separated by $S = (X_2, X_3)$.

```
solve(cor(cbind(X1,X2,X3,X4))) #almost zero in the (1,4) element

+             X1        X2        X3          X4
+ X1  5.97461670 -1.377982 -4.506875  0.06114969
+ X2 -1.37798220  4.144274 -3.410973  2.98911597
+ X3 -4.50687485 -3.410973 10.407201 -4.77295433
+ X4  0.06114969  2.989116 -4.772954  4.18905355

coef(summary(lm(X1~X4+X2+X3))) #conditioning on X2,X3 makes the coefficient for X4 zero.

+                 Estimate  Std. Error   t value      Pr(>|t|)
+ (Intercept)  0.170584034 0.005552513 30.721954 3.878611e-198
+ X4          -0.005089316 0.004164209 -1.222156  2.216773e-01
+ X2           0.162465969 0.005638442 28.813984 1.863278e-175
+ X3           0.336479362 0.004831771 69.638930  0.000000e+00

coef(summary(lm(X4~X1+X2+X3))) #conditioning on X2,X3 makes the coefficient for X1 zero.

+                Estimate  Std. Error     t value  Pr(>|t|)
+ (Intercept) -0.01699060 0.013949914   -1.217971 0.2232636
+ X1          -0.02935642 0.024020182   -1.222156 0.2216773
+ X2          -1.01083551 0.009818144 -102.955866 0.0000000
+ X3           1.02209000 0.009771908  104.594726 0.0000000
```

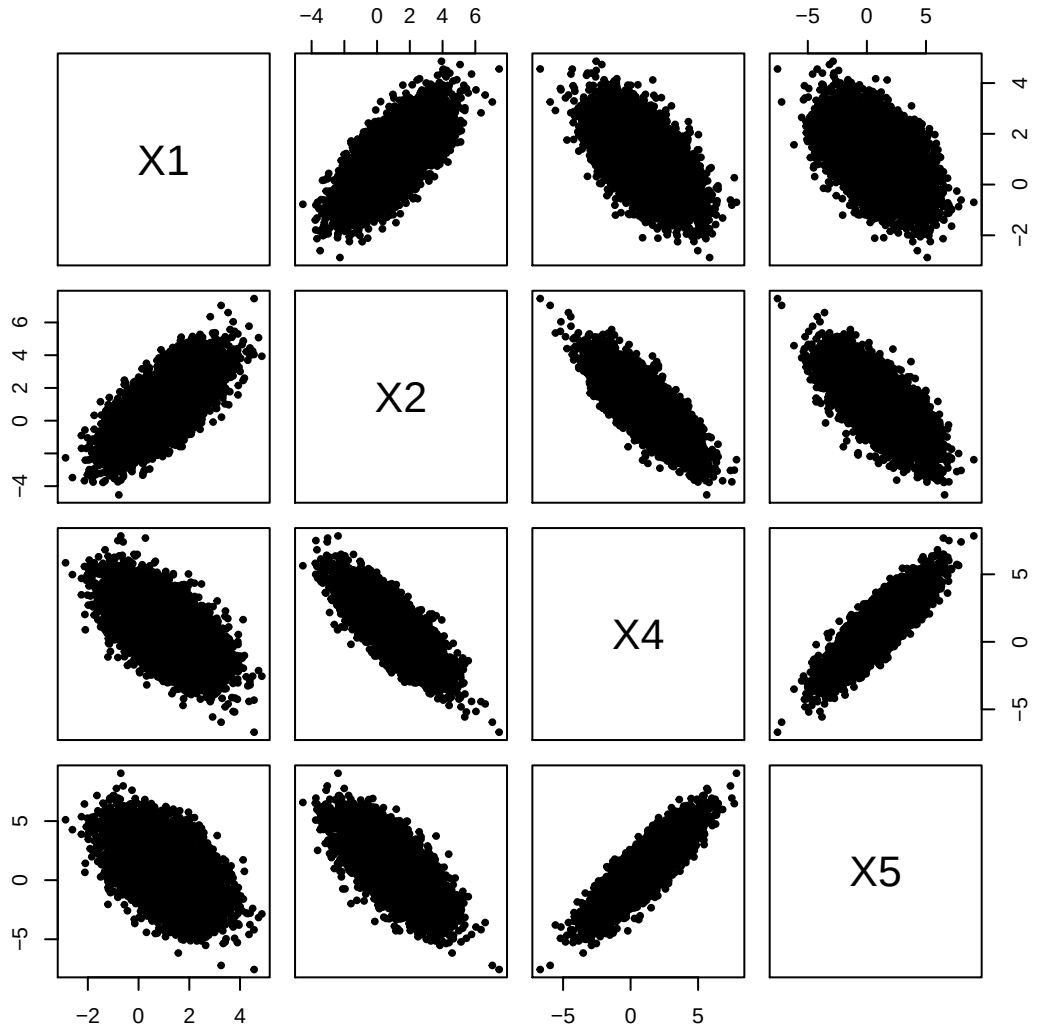
(c) **the impact of intervention:** for the intervention case we consider intervening to set $X_3$ deterministically to its marginal mean value, that is, $X_3 = 2$ with probability 1.

```
#Calculation for large N
library(mvnfast)
set.seed(22087)
n<-10000
X1<-rnorm(n,1,1)
X2<-rnorm(n,X1,1)
X3<-rep(2,n)
X4<-rnorm(n,X3-X2,1)
```

```
X5<-rnorm(n,X4,1)
X<-cbind(X1,X2,X4,X5)
cor(X)                    #Correlation

+            X1         X2         X4         X5
+ X1  1.0000000  0.7091030 -0.5775520 -0.4970122
+ X2  0.7091030  1.0000000 -0.8152654 -0.7067645
+ X4 -0.5775520 -0.8152654  1.0000000  0.8723378
+ X5 -0.4970122 -0.7067645  0.8723378  1.0000000

par(mar=c(2,2,2,2))
pairs(X,pch=19,cex=0.6)
```




```
Xb<-X
```

Note that under the intervention, the joint distribution of $(X_2, X_4, X_5)$ is now different; in the original DAG, $(X_2, X_4)$ and $(X_2, X_5)$ are marginally uncorrelated

```
cor(Xa[,c(2,4)])

+              X2           X4
+ X2  1.000000000 -0.001229615
+ X4 -0.001229615  1.000000000

cor(Xa[,c(2,5)])

+             X2          X5
+ X2 1.000000000 0.004090683
+ X5 0.004090683 1.000000000
```

whereas under the intervention, they are not.

```
cor(Xb[,c(2,3)])

+            X2         X4
+ X2  1.0000000 -0.8152654
+ X4 -0.8152654  1.0000000

cor(Xb[,c(2,4)])

+            X2         X5
+ X2  1.0000000 -0.7067645
+ X5 -0.7067645  1.0000000
```

This is due to the specific settings of the parameters used in the simulation; in general the correlation will change, but the lack of correlation in the original simulation does not hold in general, as the next simulation indicates: for example if we change the model to

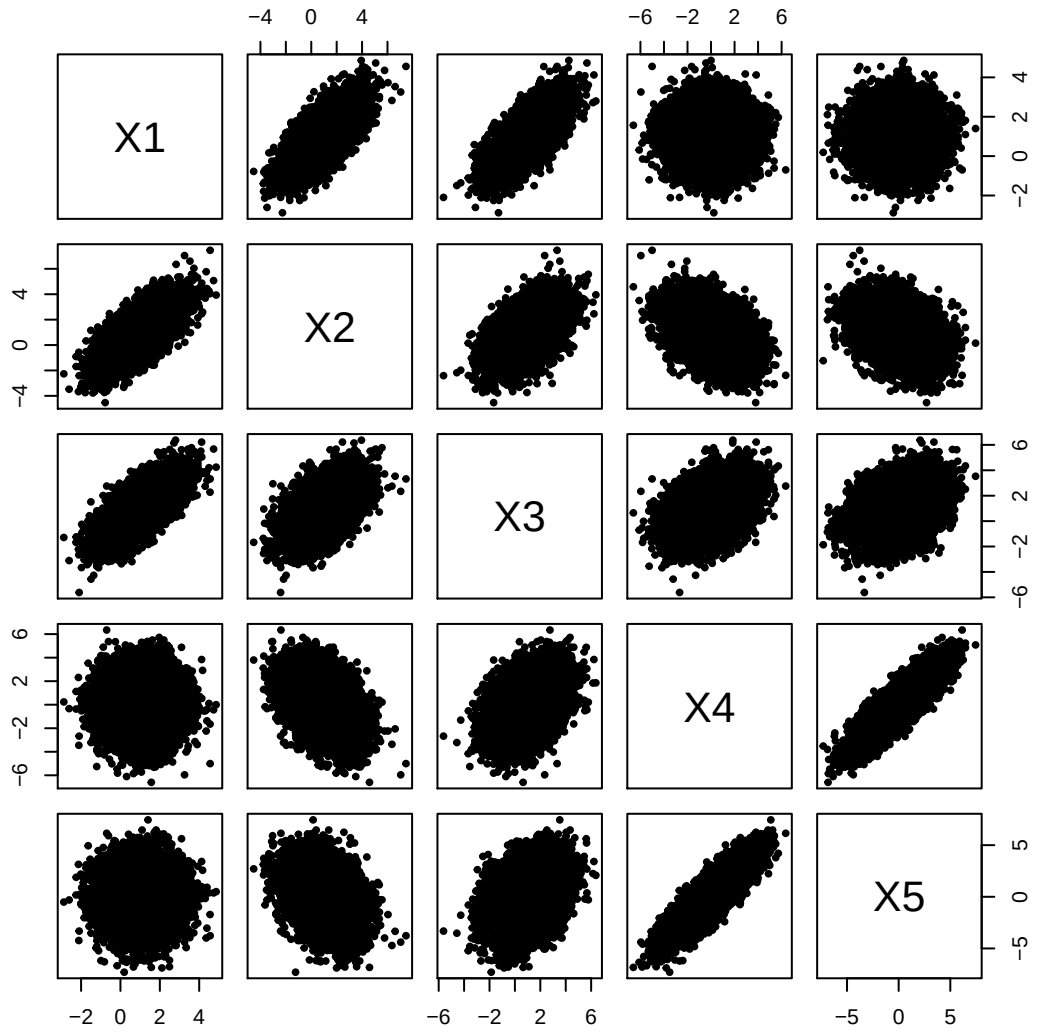
- $X_1 \sim Normal(1,1)$
- $X_2|X_1 = x_1 \sim Normal(x_1,1)$ and $X_3|X_1 = x_1 \sim Normal(x_1,1)$ independent
- $X_4|X_2 = x_2, X_3 = x_3 \sim Normal(x_3 - x_2, 1)$
- $X_5|X_4 = x_4 \sim Normal(x_4,1)$

then the correlation pattern changes.

```
#Calculation for large N
library(mvnfast)
set.seed(22087)
n<-10000
X1<-rnorm(n,1,1)
X2<-rnorm(n,X1,1)
X3<-rnorm(n,X1,1)
X4<-rnorm(n,X3-X2,1)
X5<-rnorm(n,X4,1)
X<-cbind(X1,X2,X3,X4,X5)
cor(X)                    #Correlation

+              X1         X2        X3            X4           X5
+ X1 1.0000000000  0.7091030 0.7063644  0.0005433245  0.004358544
+ X2 0.7091029770  1.0000000 0.5017080 -0.4079515182 -0.347567294
+ X3 0.7063644092  0.5017080 1.0000000  0.4168562943  0.363786116
+ X4 0.0005433245 -0.4079515 0.4168563  1.0000000000  0.866048070
+ X5 0.0043585438 -0.3475673 0.3637861  0.8660480695  1.000000000

par(mar=c(2,2,2,2))
pairs(X,pch=19,cex=0.6)
```


X1
X2
X3
X4
X5
−4 0 4
−6 −2 2 6
−2 0 2 4
−4 0 4
−6 −2 2 6
−6 −2 2 6
−2 0 2 4
−6 −2 2 6
−5 0 5