Assignment 2: Bayesian networks

Graphical and Hidden Markov Models

Ignacio Almodóvar Cárdenas & Javier Muñoz Flores

22-05-2022

Explain Dataset

We have chosen a dataset located in kaggle. It contains information about the classification of certain drug types based on different features such as the age, the sex, the blood pressure levels, the cholesterol levels and the sodium-to-potassium ratio.

The dataset contains different types of variables, including continuous and categorical. These are:

- Age: Continuous integer variable showing the age of each observation.
- Sex: Categorical variable with two levels (F,M) for indicating Female and Male respectively.
- BP: Categorical variable referring to patient's blood pressure. The levels are NORMAL and ABNORMAL. This last level includes both LOW and HIGH blood pressure conditions.
- Cholesterol: That is again a categorical variable including the cholesterol levels of the patient. It includes both NORMAL and HIGH levels.
- Na_to_K: Continuous variable showing the Sodium-potassium balance
- Drug: This will be our variable to predict. It contains two different levels (DrugX, DrugY). Depending on the characteristics of each patient it will require one drug or the other one.

```
load(file = "data.RData")
data$Age %<>% as.numeric()
data$Sex %<>% as.factor()
data$BP %<>% as.factor()
data$Drug %<>% as.factor()
data$Cholesterol %<>% as.factor()
summary(data)
```

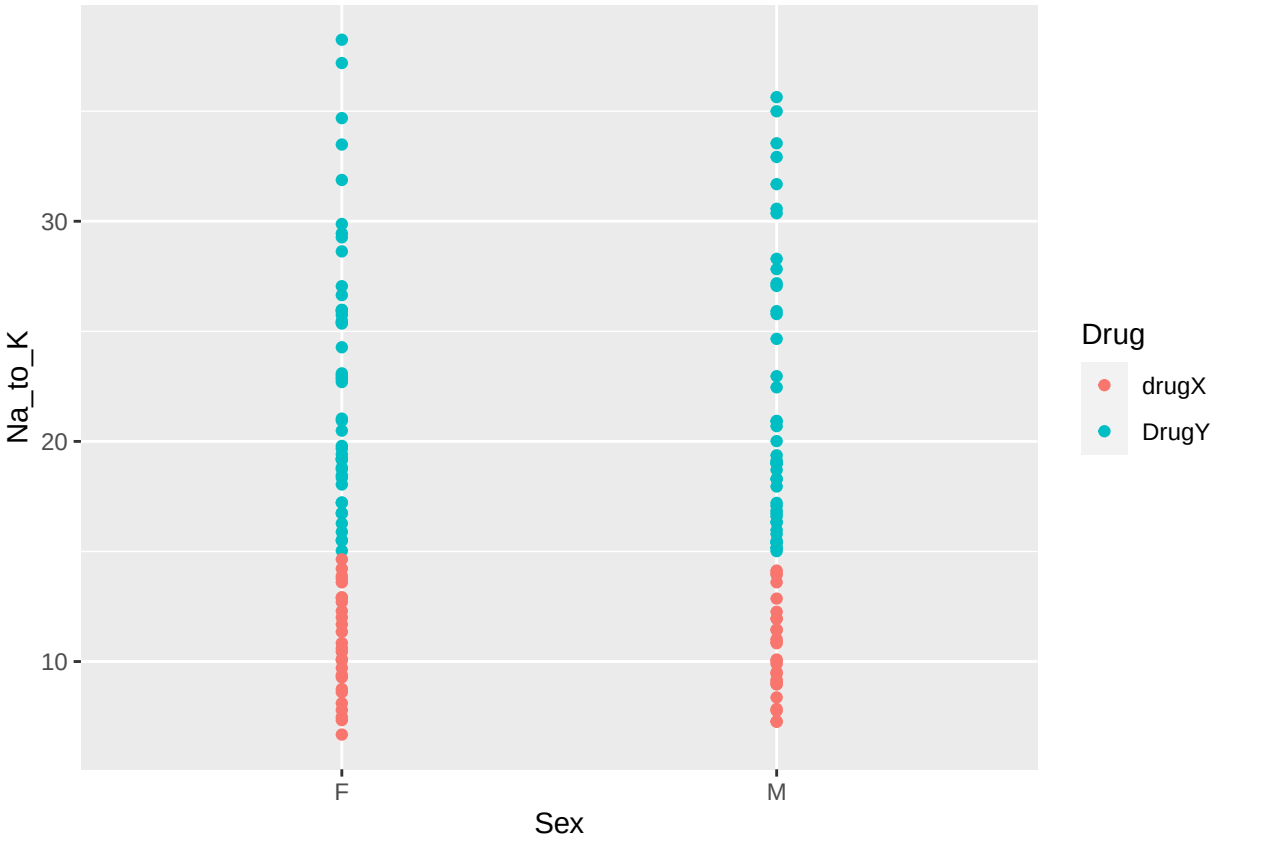
```
##       Age        Sex           BP     Cholesterol    Na_to_K          Drug
##  Min.   :15.00   F:74   ABNORMAL:86   HIGH  :67   Min.   : 6.683   drugX:54
##  1st Qu.:30.00   M:71   NORMAL  :59   NORMAL:78   1st Qu.:11.939   DrugY:91
##  Median :43.00                                    Median :16.753
##  Mean   :43.85                                    Mean   :18.009
##  3rd Qu.:58.00                                    3rd Qu.:22.905
##  Max.   :74.00                                    Max.   :38.247
```

As it can be seen, the dataset is more or less well balanced for all the categorical variables except for the Drug one, where the number of levels have a significant difference. However, as this is just an experiment to deal with Bayesian networks and we do not really care about the results obtained, the data contained should be enough to make reasonable predictions.

Define a possible network structure and give a brief explanation as to why this structure could make sense. Include a graph of the proposed model.

Initially both of us do not have much of a clue about any healthcare and the relationship between any of the existent variables. Also, the drugs given do not represent a real drug that you can check in the market or find information about it in Google. Therefore, it is difficult to find prior relationships. However, we will be looking at the data and see if we can find any relation at all.

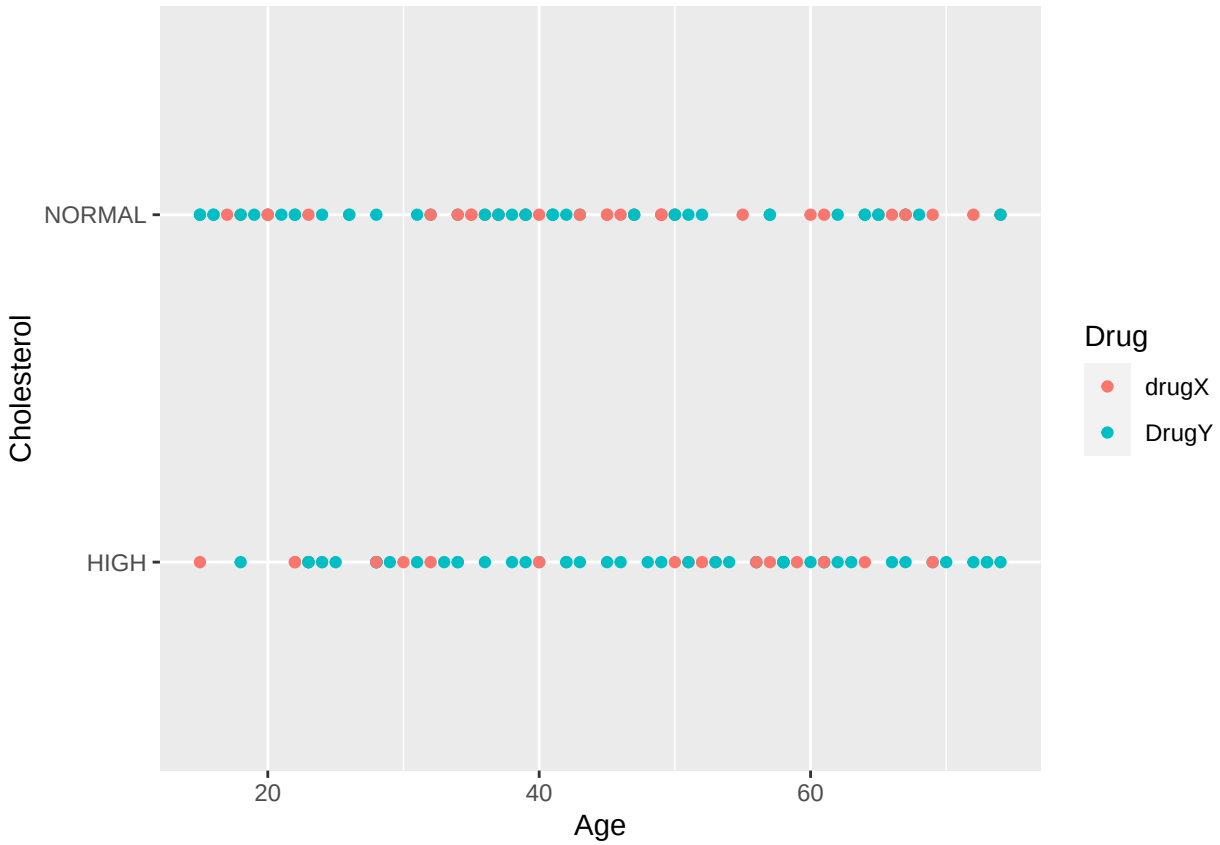
```
ggplot(data=data,aes(x=Sex,y=Na_to_K,color=Drug)) + geom_point()
```



Within this plot we are already finding some evidences of relationships. Apparently the type of drug is very relared to the level is Na_to_K in the patient. Above a concentration of 15 it will always suggest to take the DrugY, whereas below it the chosen one will be DrugX.

On the other hand, the sex does not look like it has any relationship at all with both Na_to_K and the type of drug.

```
ggplot(data=data,aes(x=Age,y=Cholesterol,color=Drug)) + geom_point()
```



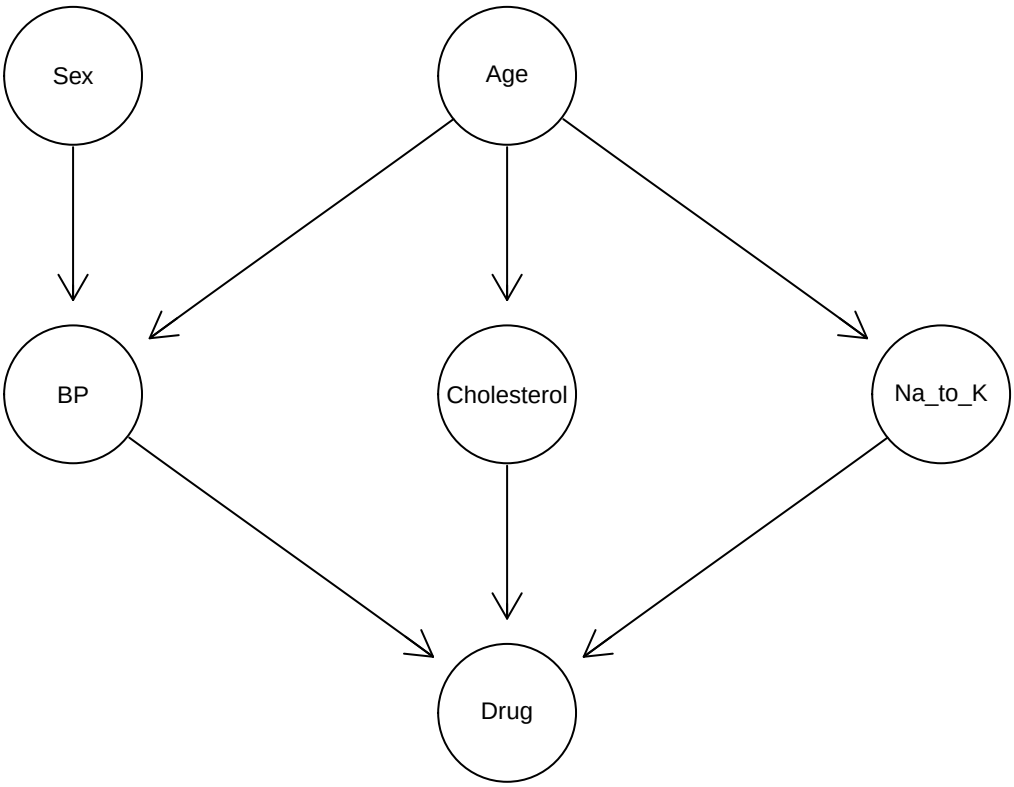
When it comes to cholesterol, it is most likely for older people to have high levels than young people. Indeed we can see that in out plot. There is not many observations for high cholesterol at ages below 30. Therefore we will assume that age and cholesterol have a dependency.

The age in general is always considered when facing medical issues. Young people tempt to have more normal levels of every aspects of health indicators. Therefore we will say that the age is indeed related to every other variable except the Sex, which has nothing to do.

We also have found some evidences on the internet that the sex can be related to blood pressure so ww will also include this deppendency on the graph.

Within this information we assume that this data might have the following structure:

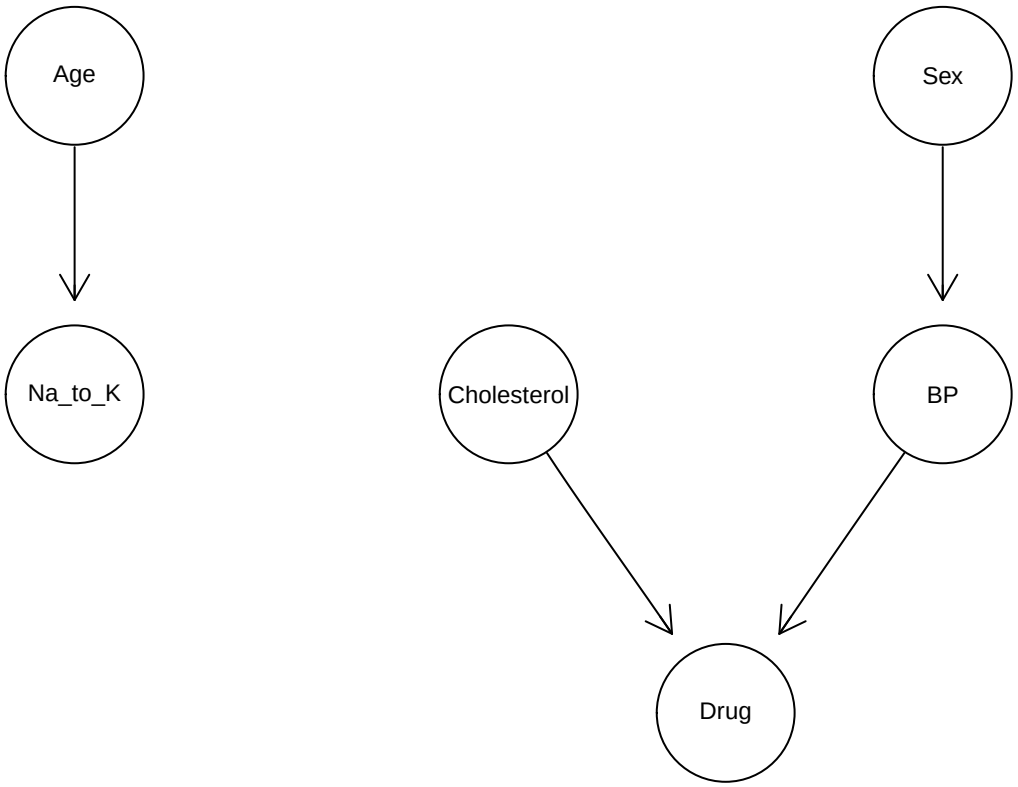
```
graph <- model2network("[Age][BP|Age:Sex][Cholesterol|Age][Na_to_K|Age][Sex][Drug|BP:Cholesterol:Na_to_
graphviz.plot(graph)
```



However, as we said before, we are dealing with some continuous variables in this case. Therefore, in order to be able to model it as a Bayesian network, we have to assume some Gaussian conditions such that any categorical node cannot have a continuous parent. Therefore we will remodel our graph to satisfy this condition and be able to pose the problem correctly.

With this said, the network will now be restructured to this one:

```
graph_2=model2network("[Age][Sex][Cholesterol][Na_to_K|Age][BP|Sex][Drug|BP:Cholesterol]")
graphviz.plot(graph_2)
```



Fit the model and use it to make an out of sample prediction

We can now easily fit by maximum likelihood the network given in the graph above.

```
mlefit <- bn.fit(graph_2,data)
mlefit
```

```
##
##   Bayesian network parameters
##
##   Parameters of node Age (Gaussian distribution)
##
## Conditional density: Age
## Coefficients:
## (Intercept)
##    43.84828
## Standard deviation of the residuals: 16.75532
##
##   Parameters of node BP (multinomial distribution)
##
## Conditional probability table:
##
##           Sex
## BP                 F         M
##   ABNORMAL 0.5945946 0.5915493
##   NORMAL   0.4054054 0.4084507
##
```

```
##   Parameters of node Cholesterol (multinomial distribution)
##
## Conditional probability table:
##      HIGH   NORMAL
## 0.462069 0.537931
##
##   Parameters of node Drug (multinomial distribution)
##
## Conditional probability table:
##
## , , Cholesterol = HIGH
##
##        BP
## Drug     ABNORMAL    NORMAL
##   drugX 0.0000000 0.5405405
##   DrugY 1.0000000 0.4594595
##
## , , Cholesterol = NORMAL
##
##        BP
## Drug     ABNORMAL    NORMAL
##   drugX 0.3214286 0.7272727
##   DrugY 0.6785714 0.2727273
##
##
##   Parameters of node Na_to_K (Gaussian distribution)
##
## Conditional density: Na_to_K | Age
## Coefficients:
## (Intercept)          Age
## 19.00257413  -0.02267038
## Standard deviation of the residuals: 7.572041
##
##   Parameters of node Sex (multinomial distribution)
##
## Conditional probability table:
##          F         M
## 0.5103448 0.4896552
```

Once we have the model fitted we can predict for example the probability for any person to take the DrugY using Bayes, which is:

$$P(DrugY) = \sum_{Cholesterol,BP} (P[Drug = DrugY|Cholesterol, BP]P(Cholesterol)P(BP))$$

This expression requires several different probabilities that we can extract from the output given above. However, there are some others that we will have to calculate, such us the probability of blood pressure (normal and abnormal) given that they come from the sex.

So first of all lets calculate the probability for normal BP.

$$P(BP = "Normal") = P(BP = "NORMAL"|Sex) =$$
$$= P(BP = "Normal"|Sex = "M")P(Sex = "M") + P(BP = "Normal"|Sex = "F")P(Sex = "F")$$

```
v <- mlefit$BP  # Try this with bayesfit or expertfit instead.
pBPgivenSex <- v$prob
v <- mlefit$Sex
pSex <- v$prob
pBP <- unname(pBPgivenSex[2,1]*pSex[1]+pBPgivenSex[2,2]*pSex[2])
pBP
```

```
## [1] 0.4068966
```

On the other hand, the probability for cholesterol can be directly obtained as it is not conditioned to any other node.

```
v=mlefit$Cholesterol
pCholesterol=v$prob
pCholesterol
```

```
##     HIGH   NORMAL
## 0.462069 0.537931
```

We can finally compute the whole probability for an individual to take the drugY using the information given in the output.

```
v <- mlefit$Drug$prob

p1=v[2,2,2]*pCholesterol[2]*pBP
p2=v[2,1,2]*pCholesterol[2]*(1-pBP)
p3=v[2,2,1]*pCholesterol[1]*pBP
p4=v[2,1,1]*pCholesterol[1]*(1-pBP)

pDrugY=unname(p1+p2+p3+p4)
pDrugY
```

```
## [1] 0.6366322
```

We can also compute for example the probability of giving a the DrugY to a normal individual, that is normal blod pressure and normal cholesterol. Using likelihood weightings we obtain:

```
cpquery(mlefit, (Drug=="DrugY"), evidence=list(BP="NORMAL",Cholesterol="NORMAL")
        ,method="lw")
```

```
## [1] 0.2698151
```

```
cpquery(mlefit, (Drug=="DrugY"), evidence=list(BP="NORMAL",Cholesterol="NORMAL")
        ,method="lw",n=10000)
```

```
## [1] 0.2784191
```

```
cpquery(mlefit, (Drug=="DrugY"), evidence=list(BP="NORMAL",Cholesterol="NORMAL")
        ,method="lw",n=100000)
```

```
## [1] 0.2725888
```

```
cpquery(mlefit, (Drug=="DrugY"), evidence=list(BP="NORMAL",Cholesterol="NORMAL")
        ,method="lw",n=1000000)
```

```
## [1] 0.2721249
```

Therefore, we can conclude by saying that when a random individual comes, without any knowledge about its health, he will be prescribed to take the drugY with a probability of 63%. However, knowing that he has normal levels of blood pressure and cholesterol, he is not very likely to take the drugY.

Fit the graph structure of the model using one or more of the different approaches we saw in class.

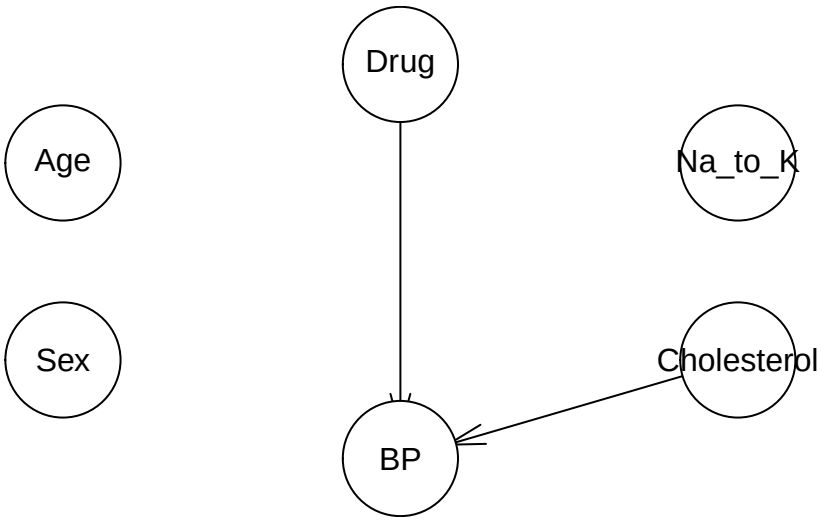
In the previous sections, we have assumed that the structure of the network were the correct one basing on different logical approaches. However, it exists different procedures to obtain the graph structure in a more unbiased way.

We will try to implement the three approaches saw in class:

- **Constraint based**: with the `aimb` method.
- **Score based**: with the `hc`method. We will use both AIC and BIC score functions to evaluate the fit.
- **Hybrid**: with the `mmhc` method.

First, we perform the first approach mentioned. This technique can be slow as it has to analyse each pair of nodes in the network. However, since our network is not too very large, the computational cost is not very high.

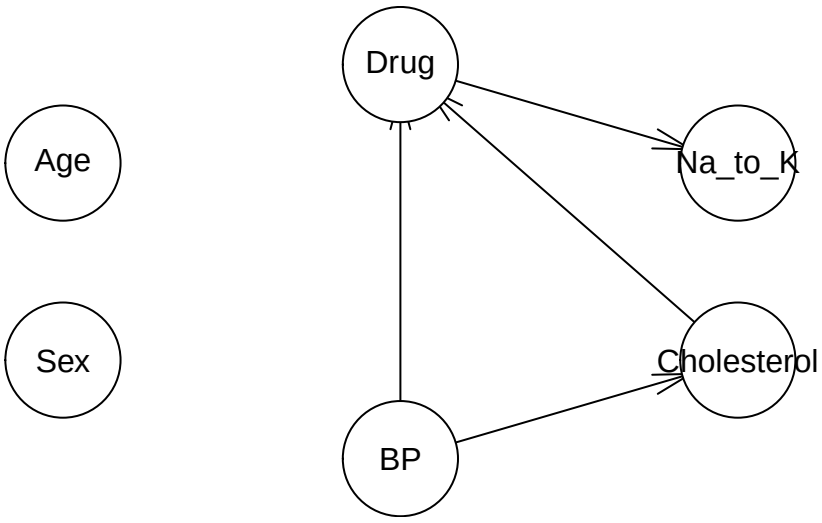
```
structure_iamb <- iamb(data)
plot(structure_iamb)
```



The structure obtained by this technique suggested that three out of the six nodes have not dependencies with others. This composition does not agree with our first approach, since it does not take into account some relationships we have assumed that were important and it eliminates so many dependencies.

Let's analyse the structure suggested by the **score based** methods. First, we look at the output basing on BIC.

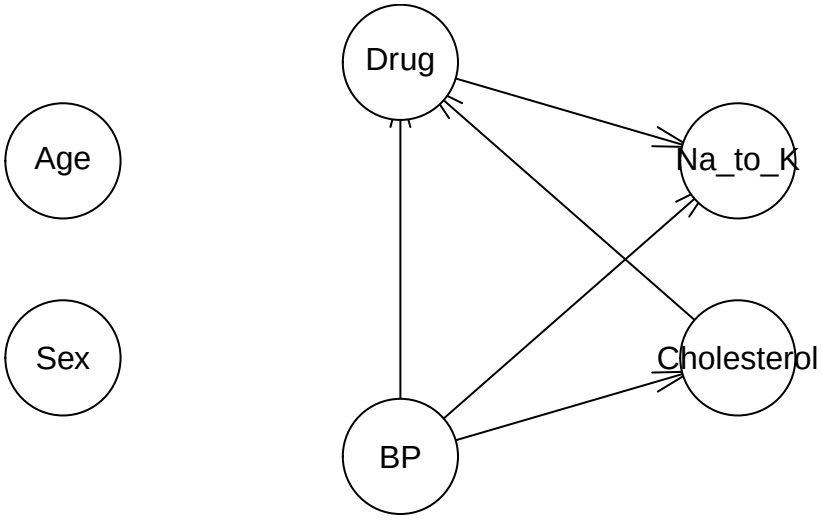
```
structure_bic <- hc(data, score = "bic-cg")
plot(structure_bic)
```



As we can see in the plot, this method includes the *Cholesterol-Drug* and *Drug-Na_to_K* dependencies. While the first of the two relationships could makes sense, the second one does not make it. It could make sense an opposite direction of the arrow between *Drug-Na_to_K*, but in Conditional Gaussian Bayesian networks like this, categorical nodes have no continuous node parents, so it cannot be possible.

Let's see the structure proposed using AIC method.

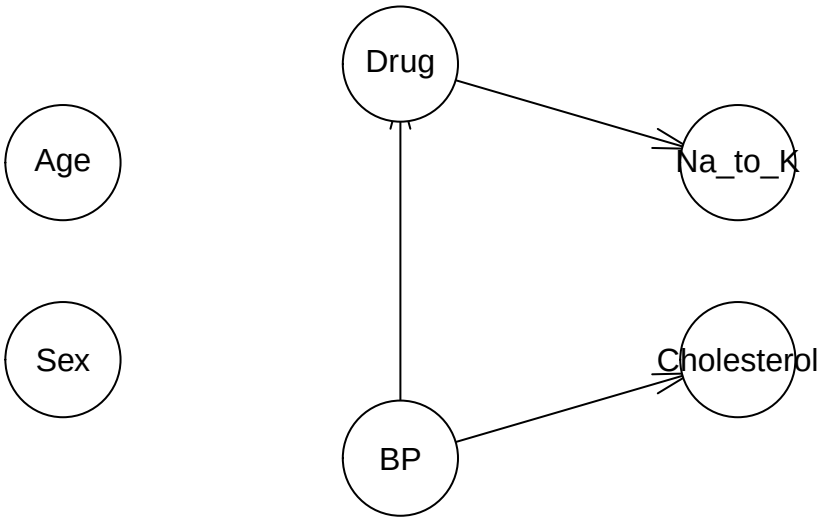
```
structure_aic <- hc(data, score = "aic-cg")
plot(structure_aic)
```



We can see that the dependency *BP-Na_to_K* have been added, which is reasonable. Nevertheless, the dependency *Drug-Na_to_K* stays.

Finally, we will see the struture suggested by the last of the approaches mentioned: **hybrid method**, a combination of the two previous ones.

```
structure_mmhc <- mmhc(data)
plot(structure_mmhc)
```



Again, it considers *Drug-Na_to_K* dependency and it does not include important relationships which we have found through the score based methods.

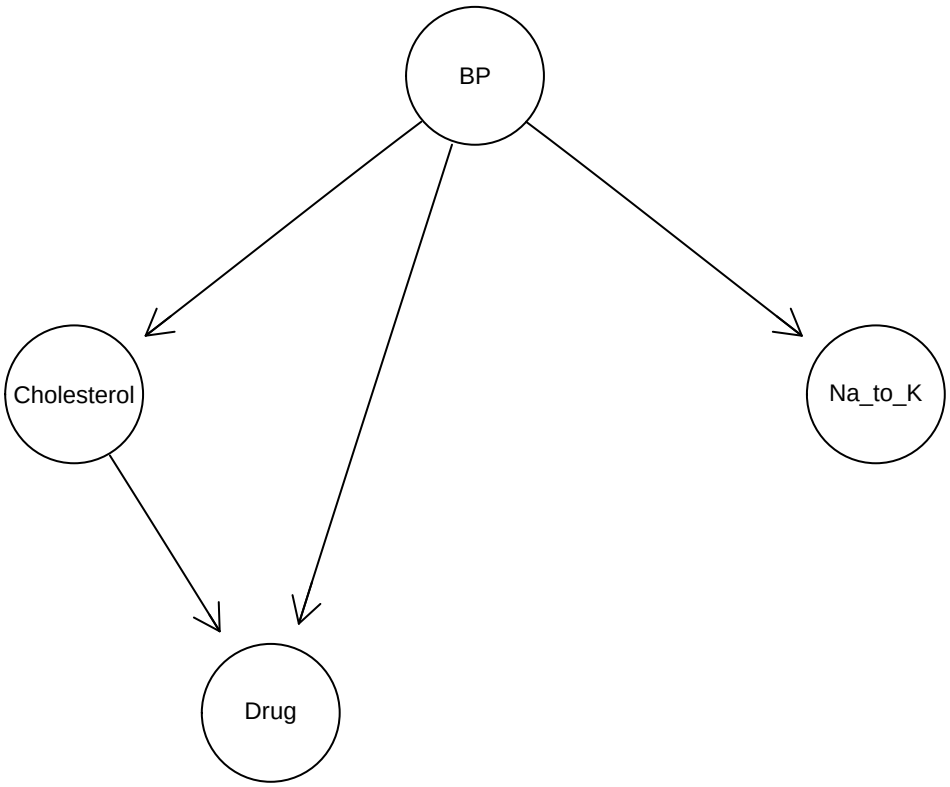
Which of the proposed graphical structures is the best?

In view of the results obtained in the previous section, where we have used algorithms which have helped us to learn the conditional independence structure, we can draw the following conclusions:

- We should eliminate the *Drug-Na_to_K* since we consider it does not make sense with the analysis of or network.
- The nodes *Age* and *Sex* should be eliminated since it does not appear in any of the four structures proposed by the methods.
- The structure suggested by the score based methods adds important relationships which should be considered from a health point of view.
- The structure proposed by the AIC method should be selected as the best one instead of the one suggested by the BIC method because it considers *BP-Na_to_K* dependency. According to some research, the sodium (Na)/potassium (K) ratio could be associated with blood pressure (BP), so this relationship should be added to our network (Reference).

Thus, we delete the *Drug-Na_to_K* dependency as well as the nodes *Age* and *Sex* and we draw the most optimal graphical structure.

```
final_structure <- model2network("[BP][Na_to_K|BP][Cholesterol|BP][Drug|BP:Cholesterol]")
graphviz.plot(final_structure)
```



Summary of the results.

To sum up, we conclude this work enumerating the most important insights obtained:

- The first approach for modelling the graph structure, i.e. the structure assumed of ourselves, has not been matched with the ones proposed by the algorithms. However, the dependency *Cholesterol-Drug*, which stays in almost all the structures proposed, seems to be the most reliable.
- It has been very important the assumptions in a network with both categorical and continuous variables like ours, i.e. a Conditional Gaussian Bayesian network. It is because we have taken into account that categorical nodes have no continuous node parents and it has not allowed to draw some dependencies that at first we could have considered to include.
- The probability that a random patient whose health we know nothing about, is prescribed to take the drugY is around 63%. Nevertheless, if we know that a patient has normal levels of blood pressure and cholesterol, it is more likely that he should take the drugY.
- The algorithms that provide a network structure agree that the variables *Age* and *Sex* should not be included in the network.