Lab 1 report

Probabilistic Graphical Models

Authors:
Hussein AWALA
Amine BIROUK
Bonpagna KANN
Assem SADEK

Supervisors:
Prof. Jean-Baptiste DURAND
Prof. Xavier Alameda-Pineda

December 15, 2018

Contents

1 Introduction

Probabilistic graphical model estimation, is one of the on going research topic. The aim for the estimation is to find a mapping graph that can represent the relations between different distributions, given a sample data.

In this report, we will answer the questions on probabilistic graphical models lab to compare between different methods for estimation the mapping Graph.

2 Answers

2.1 Lab Work

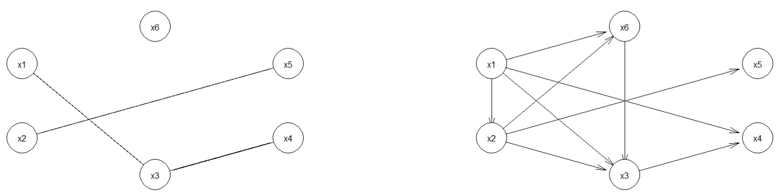
2.1.1 Simulated Data

Several comments must be noted on the obtained results:

1. Figures 1, 2 and 3 shows that Hill climbing (HC) outperformed Gaussian S (GS) in every iteration. HC try to construct a directed graph (DAG), while GS can give a partial directed graph (PDAG). GS focus on getting the relationships between the variables, and this leads to get a good skeleton but with less information about the direction of the edge.

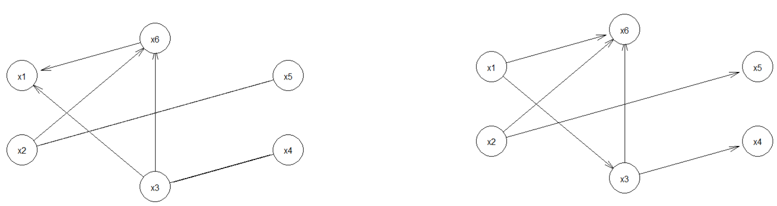
2. GS failed to construct the required graph, even when the procedure ran over 1000 data sample.

3. Figure 2b shows that HC has already got his maximum estimation for mapping for the data from iteration 60. There's no difference between this graph and the graph in figure 3b



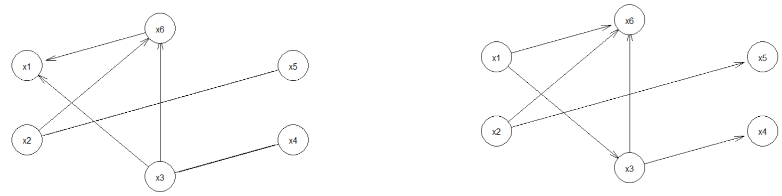
(a) gs graph estimation (b) hc graph estimation

Figure 1: The estimation using 40 examples of the dataset



(a) gs graph estimation (b) hc graph estimation

Figure 2: The estimation using 60 examples of the dataset



(a) gs graph estimation

(b) hc graph estimation

Figure 3: The estimation using 100 examples of the dataset

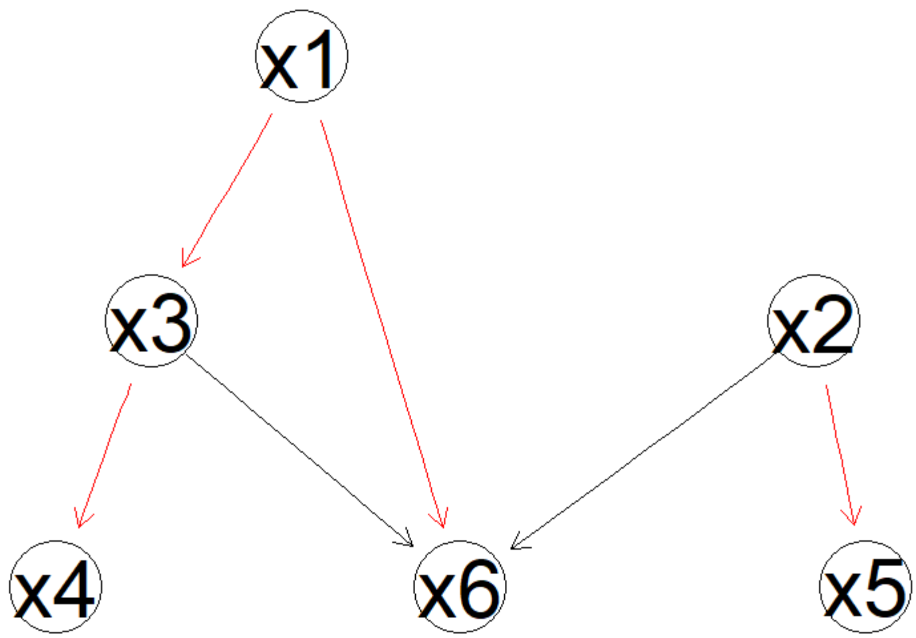


Figure 4: Difference between the two generated graph given on the 60 examples dataset

2.1.2 Real data: asset returns

1. During creating the data frame, empty spaces in the CSV file had to be removed, and the NA value, found in the data frame, had to be removed too.

2. Figures 5 and 6 show the constructed graph using gs and hc procedures respectively.

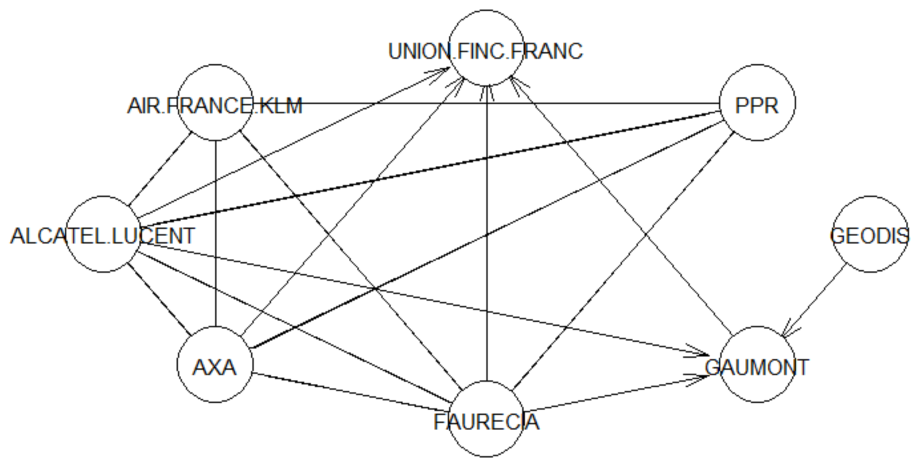


Figure 5: Graph generated by GS on Asset returns data

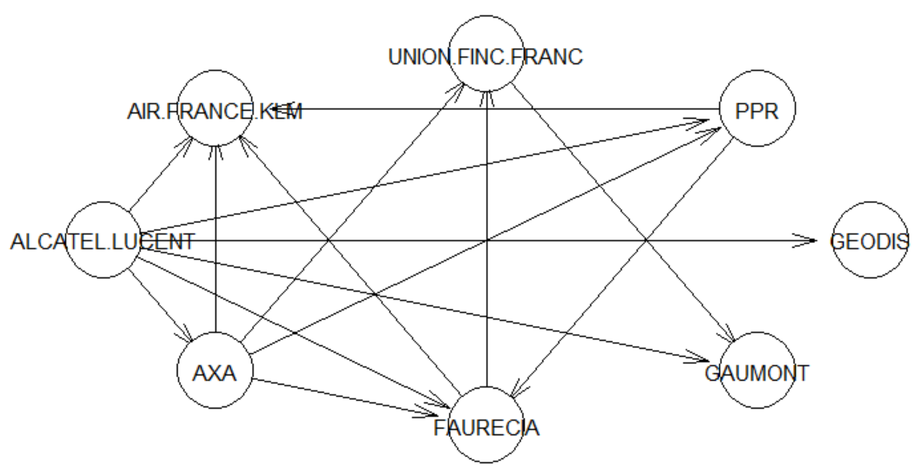


Figure 6: Graph generated by HC on Asset returns data

3. According to Table 1, we choose two marginal dependencies found by GS but not with HC. Both of them have p-value very low (less than $\alpha = 0.05$), this implies that we will reject the two hypothesis that claim : GEODIS and ALCATEL are independent and GEODIS and PPR are independent. HC has stronger estimation in these two cases.

Found by GS but not by HC	
Relation	p-value
$GEODIS \perp\!\!\!\perp ALCATEL$	4.625×10^{-5}
$GEODIS \perp\!\!\!\perp PPR$	1.381×10^{-4}

Table 1: ci.test on marginal independencies found by GS but not by HC

4. According to Table 2, we choose two conditionals dependencies found by HC but not with GS. Both of them have p-value very low (less than $\alpha = 0.05$), this implies that we will reject the two hypothesis, This time GS has stronger interpretation on the relation between these variables.

Found by HC but not by GS	
Relation	p-value
$GEODIS \perp\!\!\!\perp GAUMONT \mid ALCATEL$	3.783×10^{-3}
$ALCATEL \perp\!\!\!\perp UNION \mid AXA, FAURCIA$	3.707×10^{-3}

Table 2: ci.test on conditional independencies found by HC but not by GS

5. In this case both procedures have agreed that GEODIS and UNION are independent given GAUMONT and ALCATEL. The p-value in table 3 is much larger than $alpha = 0.05$. This time the null hypothesis is accepted which means that in fact both variables are independent given the mentioned conditions. This was the most probable result, since both graph has already agreed on it which push forward that such conditional independency would be valid.

Found by GS and HC	
Relation	p-value
$GEODIS \perp\!\!\!\perp UNION \mid GAUMONT, ALCATEL$	0.1098

Table 3: ci.test on conditional independencies found by GS and HC

2.2 Mandatory additional questions

Question 1 : Defining Consistency and Choosing an algorithm

We define the consistency of directed PGM estimator algorithm as follow: If the sample size m tends to ∞, the resultant mapping graph G tends to G^*. By definition G^* is the perfect map/representation of the distribution P of the random variables V in the dataset. So we note it like :

$$\lim_{m \to \infty} G = G^* \tag{1}$$

Equivalent to,

$$\lim_{m \to \infty} I(G) = I(P) \tag{2}$$

Where I(*) denotes the set of independent relationships.

Choosing an estimator algorithm: Min-Max Hill Climbing (MMHC)

Before choosing the algorithm, some research has been done on available literature survey [1], we found that the PGM estimation algorithms are categorized in three main approaches : The first one like GS, which is a constraint-based approach algorithm, try to identify the set of conditional dependencies in the distribution P, then try to construct the graph to satisfy these constraints, this leads to a global structure graph that normally is a PDAG.The main problem of this kind of algorithm is that they lack a sort of objective function to measure how good a graph is. This deprive of doing optimization on the output structure. The second approach is like the HC algorithm, a score-based approach algorithm. The approach try to define a score function, which help to indicate how well the graph maps the data. To get a good score, heuristic methods is used, like greedy search to find the one that optimize the value of the objective function. The main drawbacks of this approach is the computation and it can be stuck in a sub-optimal structure. The third approach is the regression-based approach, which is a recent approach that put the variables in a regression framework (e.g. linear regression), which essentially guarantee to find a global optimization for the objective function.

The chosen algorithm is Min-Max Hill Climbing (MMHC). this algorithm was introduced in 2006 by Tasmardinos et al. [2]. This algorithm can be categorized as a hybrid approach which means that it combines between more than one of the mentioned approaches. In this case, it combines between the constrained-based and score-based approaches. It has outperformed several of the state of the art algorithms, like Sparse Candidate algorithm (SC). The algorithm is divided in two procedures, one focus on getting the skeleton of the graph and the other focus on find the right orientation of the edges.

The first procedure uses the Max-Min parents and children algorithm (a special version published by the same author in 2003). According to the most commun version explained by Mikael Petersson [3], the MMHC try to find a skeleton of the bayesian network before going to the second procedure. Thus, it doesn't decide the orientation of the arcs, it only use the constrained-based approach by inserting and edge (X,Y) if and only if $X \perp\!\!\!\perp Y \mid SforanysubetS$. The algorithm is divided in three main stages: Stage 1 focus on getting the parents of a variable X.Stage 2 focus on getting the children of the same variable. Stage 3 acts as a symmetry correction to remove any duplication.

The second procedure uses the greedy hill-climbing search like HC procedure, it's score-based method. It takes the output of the first procedure as an input. This means that using only the edges found in the skeleton, it try to modify in the edge between two variables by doing the operation of adding and removing the arc by changing the orientation without producing a previously visited graph structure. Keeping doing this operation while saving the graph with the high score and return it at the end.

We have choose this algorithm for many reasons. First, it combines between the approach, thus we can benefit from the features of each approach simultaneously. Second, it could be considered as a evolution for HC procedure, because it uses the same hill climbing algorithm but it adds on the top of it another method from the constrained-based approach. The last reason came from the paper's evaluation sections which prove that the algorithm has beaten many of the start of the art algorithm at the moment and also considered as a computational efficient method on the large scale data with many variables.

Question 2 : Compare MMHC to HC

1. Figure 7 shows the constructed graph using mmhc procedure respectively.

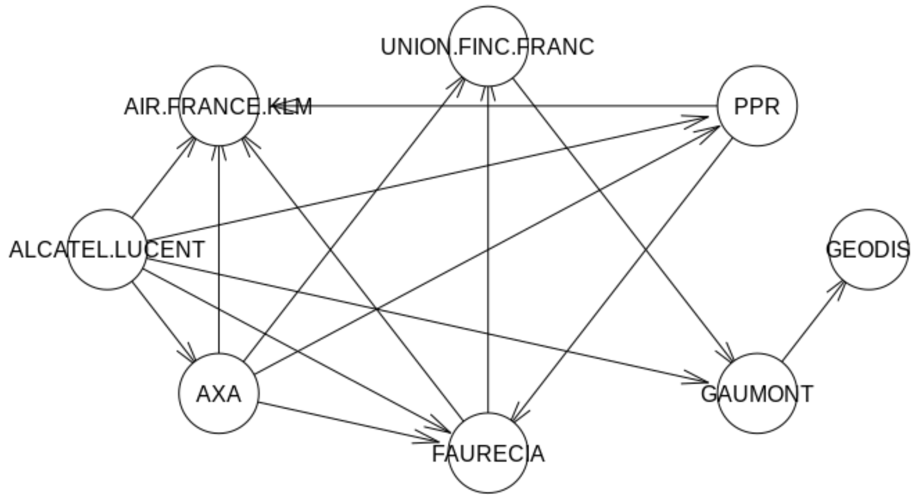


Figure 7: Graph generated by MMHC on Asset returns data

2. There's no difference in marginal independencies between the two graphs.
3. According to Table 2, we choose the only conditional dependencies found by HC but not with MMHC. The p-value very low (less than $\alpha = 0.05$). this implies that we will reject the hypothesis again and assuming that the right hypothesis is in MMHC.

Found by HC but not by MMHC	
Relation	p-value
$GEODIS \perp\!\!\!\perp GAUMONT \mid ALCATEL$	3.783×10^{-3}

Table 4: ci.test on conditional independencies found by HC but not by MMHC

4. The two methods have nearly the same structure of the graph. They share a lot of commun conditional independencies. They both agree on the conditional independency in table 3. We can conclude that this relationship has been accepted not only by GS and HC, but also by MMHC. This could help us to be more sure about the validity of the relationship between GEODIS and UNION.

More general comment: In both scenario, we found that HC has stronger estimation over the MMHC. This leads us to search for a score metric to evaluate both methods aiming to find a decision over which is better over this dataset. Since both of them estimate a Bayesian Network. We use the function *Score* to calculate the score of each Graph. We got:

- Score(HC) = 93713.31
- Score(MMHC) = 93709.34

We have to mention a note given on the function *Score* in the official documentation of bnlearn package that higher score values are better, because the value is scale by -2 compared to BIC and AIC scores.

According to the above result, we can see that HC is better than MMHC, but with a small difference and this is normal because when we compare the two graph visually, we will see that both of them are very similar. They differ only in the two relationships given in the tables 4 and 5. The two relationships vary between the three variables ALCATEL, GAUMONT and GEODIS. We can observe that the p-value in table 4 is higher than the one in 5, then we are more confident about the result the relationship given by HC more than MMHC.

Found by MMHC but not HC	
Relation	p-value
$GEODIS \perp\!\!\!\perp ALCATEL \mid GAUMONT$	1.064×10^{-4}

Table 5: ci.test on conditional independencies found by MMHC and not by HC

2.3 Optional additional questions

Inspired from the previous section's last part, about calculating the score for the methods who gives a full directed PGM. We suggests a simple evaluation protocol to check the consistency of an algorithm that is based on this score, since the score indicate how good we approach to the perfect map G^*. Thus, according to the giving definition of consistency 1 we can check by the score as follow :

1. Iterate over a data multiple time, and each time construct a graph using a sample size that increase at each iteration constantly (e.g. $m_1 = 500, m_2 = 1000, ..., m_T = m$, where m = the full size of the data and T = the number of iterations.)
2. At each step, calculate the score of the graph relative to the chosen sample data.
3. Plot at the end, a graph that shows the size of data at each iteration Vs. the score.
4. If we got an positive correlated relationship between the size and the score, that should tends to a constant when m tends to ∞, this means that the algorithm is consistent.

The code is available in file q3.R

The figures 8 and 9 shows the consistency of the two methods of HC and MMHC on the asset data.

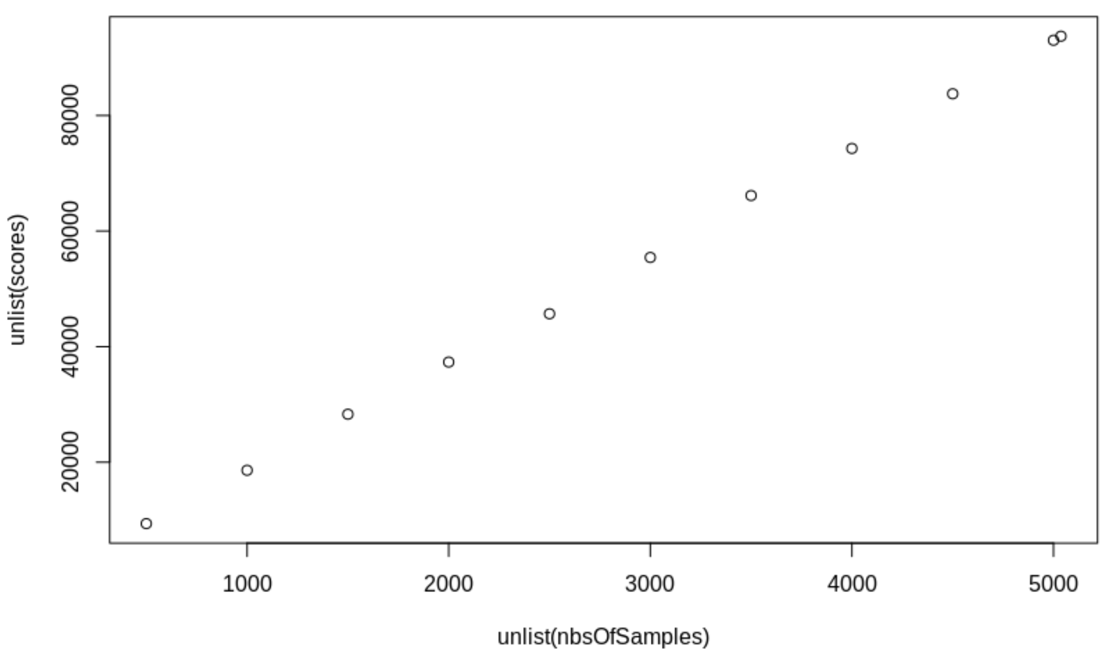


Figure 8: The plotting of HC procedure, produced by the explained protocol in section 2.3

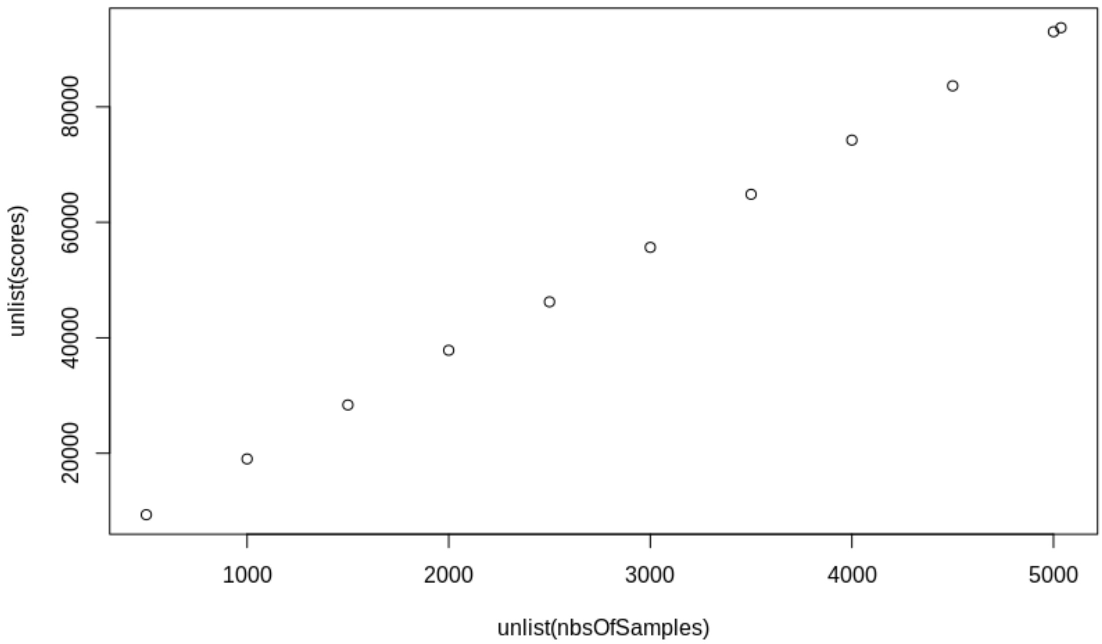


Figure 9: The plotting of MMHC procedure, produced by the explained protocol in section 2.3

List of Figures

List of Tables

References

[1] Y. Zhou (2007) *Structure Learning of ProbabilisticGraphical Models: A Comprehensive Survey.*

[2] I. Tsamardinos et al. (2006) *The max-min hill-climbing Bayesian network.*

[3] M. Petersson (2010) *The Maximum Minimum Parents and Children Algorithm.*