
**Budapest University of Technology and Economics**
Faculty of Electrical Engineering and Informatics
Department of Measurement and Information Systems

Máté Csőke

# Investigation of a causal discovery algorithm using hidden compact representations

Consultant:
Dr. Gábor Hullám

**Budapest, 2019**

# Contents

# List of Figures

# List of Tables

# 1 List of frequently used symbols and notations

$G$ = Graph
$V$ = Vertices
$E$ = Edges
$DAGG$ = Directed acyclic graph $G$
$N$ = Number of variables, or columns in data table
$X_i$ = Stochastic variables
$\{(x_i, x_j)\}$ = Ordered pair of $x_i$, $x_j$
$P(X)$ = Probability distribution of stochastic variable $X$
$P(X|Y)$ = Probability distribution of stochastic variable $X$ given $Y$
$|n|$ = Sample size
$X \to Y$ = X causes Y
$A \perp\!\!\!\perp B$ = A and B are independent stochastic variables
$A \not\perp\!\!\!\perp B|C$ = $A$ and $B$ are not independent stochastic variables given variable $C$
$c$ = Degrees of freedom
$O$ = Observed value
$E$ = Expected value
$\chi^2$ = Chi square score
$M_R$ = Row marginal
$M_C$ = Column marginal
$T$ = table of data
$\phi$ = Scoring function
$B_N$ = Set of Bayesian networks with $N$ variables

# 2 Absztrakt

A kauzális kapcsolatok elemzése számos tudományterületen lényeges szerepet tölt be, legtöbbször elengedhetetlen látni a vizsgált változók egymásra gyakorolt befolyásoló hatását, ok-okozati összefüggéseit. Azonban olyan területeken, ahol csak megfigyelési adatok állnak rendelkezésre ez jelentős kihívást jelent, ugyanakkor a tárgyterületre jellemző háttértudás ezen képes enyhíteni.

Gyakorlati alkalmazásokra számos területen találunk példát, úgymint szociológia, pénzügyi befektetések, élettudományok, azóta, hogy Judea Pearl az oksági kapcsolatok feltárásának alapjait lefektette. Ezek közül talán legszembetűnőbb az orvostudományban történő felhasználása ezen módszereknek, ahol is, a randomizált klinikai vizsgálatok nehézségeit leküzdendő jelentek meg. A Mendeli randomizációként nevezett oksági összefüggések feltárásán alapuló kutatás módszertan betegségek illetve kockázati tényezők közt térképez fel kauzális kapcsolatokat genetikai variánsok vizsgálatán alapuló modellek segítségével.

Az oksági kapcsolatok feltárására számos algoritmus jött létre az elmúlt pár évtizedben, melyek döntően két csoportra bonthatók: kényszer illetve pontszám alapú módszerekre. Kényszer alapú algoritmusoknál a háttértudás bizonyosságaira, illetve a feltételes függőségek sajátosságaira támaszkodunk egy-egy kauzális struktúra felderítése során, míg pontszám alapú algoritmusoknál egy pontszám függvényre, amely igyekszik megállapítani, hogy egyes oksági gráfok milyen jól jellemzik a rendelkezésünkre álló adatot. Mindkét megközelítés módszerei között vannak folytonos illetve diszkrét változókkal dolgozó algoritmusok. Ez utóbbiak közé tartozik a rejtett kompakt reprezentációk segítségével kauzális kapcsolatot feltáró algoritmus, melynek érdekessége hogy a számos oksági hálót jellemezni képes pontszámfüggvény közül a Schwarz-féle bayesi információs kritériumot (BIC) alkalmazza.

A dolgozatomban bemutatott kutatás célja a pontszámfüggvények és a velük járó feltételezések elemzése, illetve a rejtett kompakt reprezentációkon alapuló oksági feltáró algoritmus teljesítményének vizsgálata különböző pontszám függvények alkalmazása mellett.

# 3 Abstract

The analysis of causal relationships plays an important role in many fields of science, in most cases it is essential to see the influence of the examined variables on each other and their causal relationships. However, in areas where only observational data is available, this is a significant challenge, although background knowledge specific to the subject area can mitigate this.
Practical applications have been found in many areas, such as sociology, financial investment, life sciences, since Judea Pearl laid the foundations for exploring causation. Perhaps the most striking of these applications is the use of these methods in medicine, where they have emerged to overcome the difficulties of randomized clinical trials. The research methodology based on the discovery of causal relationships, called Mendelian randomization, explores causal relationships between diseases and risk factors using models based on genetic variants.
A number of algorithms have been developed over the past few years to uncover causal relationships which are mainly divided into two groups: constraint-based and score-based methods. Constraint-based algorithms rely on the available background knowledge and the properties of conditional dependencies in exploring a causal structure, while score-based algorithms rely on a score function to determine how well the investigated causal graphs represent the available data. Both approaches include algorithms that work with either continuous or discrete variables. The latter includes a causal relationship discovery algorithm using hidden compact representations, which is interesting due to its use of Schwarz's Bayesian Information Criterion (BIC), a scoring function that can describe many causal networks.
The aim of my dissertation is to analyze score functions and their associated assumptions, and to investigate the performance of a causal exploration algorithm based on hidden compact representations using different score functions.

# 4 Introduction

In today's world of intelligent systems research, the main emphasis lies on making predictions, and facilitating classification based on observational data. There is less focus on not predictive but descriptive modelling, especially on modelling the causal mechanisms based on observational data. As a remedy, the field of casual modelling, namely the technique of Bayesian networks can be applied. In many fields precise modelling of causal structures are indispensable, in the field of medicine.
To gain practical knowledge about causal structures through active intervention in most of the cases is simply not possible, either because of ethical reasons, or because of financial ones. The conduction of those trials is not always possible, but gathering huge amounts of observational data, in which there are no interventions, is manageable. Bayesian network based methods are one of the few approaches that allow the modelling of the underlying causal structures, which is generally a hard task, but not impossible. There are different approaches within this group of methods, and there are several types of algorithms, which enable at least a partial solution to the problem of discovering causal relationships. The conditional independence relations between variables are exploitable for this cause, just as well as there are score based methods to evaluate the fitting of local or global models to the data. These fields exist as their own, but in some cases, the different techniques could be used together as better tools for a particular problem.
The modeling of causal directions between discrete, more exactly categorical variables, was always of great difficulty, as some assumptions does not always hold for ordinary algorithms, mainly regarding the ordering.
The method of Hidden compact representations (HCR) Cai et al. 2018 tries to provide a method for specifically testing this case, with increased accuracy compared to the other algorithms. HCR builds upon both approaches, which enables performance testing on different types of score functions.
In the following sections this algorithm will be tested, through the methods of the original paper (i.e. in which the algorithm was published). In addition, the difference of the scoring functions will be investigated on both synthetic and real world datasets. The conclusion of this study is presented in the last section, which states that the originally provided scoring function for HCR is not the best performing one, also proving that the method itself has unmentioned disadvantages in cases where there is no relationship between variables. This means that without the aid of other algorithms, its ability to infer a whole causal graph is exceptionally limited.

# 5 Theoretical background

Intelligent data analysis, a union of data science and artificial intelligence methods is a popular research field which can be divided into several sub-fields. This grouping can be achieved in various ways, for example the different algorithms can be categorized according to their main approach. Classification and regression methods can be viewed as predictive methods requiring a labeled data set to perform supervised learning. Clustering and association rule mining on the other hand can be considered as descriptive methods, as they rely on observational data an perform unsupervised learning. These categories and corresponding methods can be seen on Figure 1. The structure learning of Bayesian networks, can be both predictive and descriptive depending on the problem and the available background knowledge. In cases where there exists a validated model, this model can be applied on the observed data and predictions can be made.

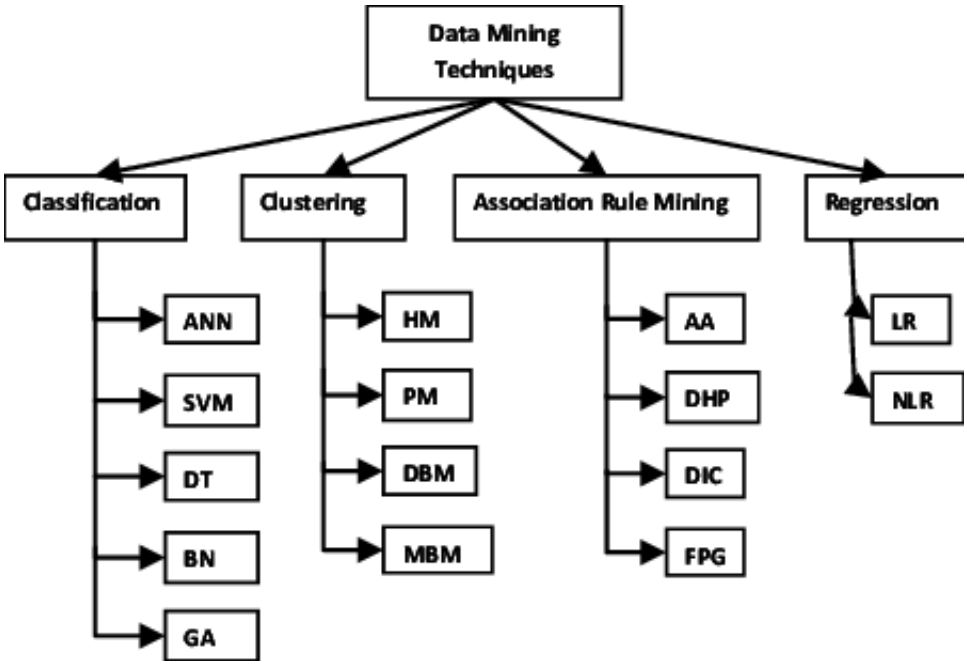



Figure 1: Fields of Data Science, Source: H. Patel and D. Patel 2014

In other cases there exists no such a model, and the goal is to learn the dependency relationships between variables, i.e. the structure of the model. In some sense, it is similar to rule mining, however in this case the relationships are conditional dependency relationships not associations. To identify relationships effectively, there has to be an understanding of the underlying model, its assumptions, specific architectural attributes that come from its stochastic nature, and ways to compare different architectures on the same datasets.

## 5.1 Bayesian networks

Since this thesis intends to incrementally improve the study of a certain type of graphical models, it is necessary to present a thorough description of the models themselves: more specifically its architecture and its descriptive capabilities.
Bayesian networks as seen on Figure 2 are models consisting of a (1) directed acyclic graph (DAG), where nodes represent stochastic observed variables, and edges describe the dependence relationship between the nodes; and a (2) conditional probability distribution that defines the conditional probability of each node given its parents. The networks themselves could be described as a structure, understood as semantics in a few different ways of approach, each with more assumptions. The order corresponding to representational capabilities from weak towards strong is as follows.

(1) From the field of information theory a network of this type, is an effective encoding of conditional independence statements. This can be seen through the following, there is a table that has $K$ dimensions, each variable taken as a boolean, the table of the full joint distribution would require $2^K$ values. While a Bayesian network in which each node is influenced by at most $M$ other variables it would take $K * 2^M$, where for each table $M$ is has much lower $K$. In a context where there is a table of 15 variables, the full joints distribution would require $2^{15} = 32768$ numbers, while a Bayesian network where each node has at most 3 parents it is $15 * 2^3 = 120$ ( Russell and Norvig 2009) .
(2) Apart from that, from the statistical point of view. The Bayesian network can be seen as a representation of a joint probability distribution over $X$.

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(x_i | Parents(X_i))$$

Each edge represents a conditional dependence between the nodes, while the graph itself is a map of the dependences in the learned joint distribution.
(3) The third interpretation with the strongest presumptions is the causal interpretation, where each edge represents a causal dependence between the variables. (András, Hullám, and Antal 2019).
Formally, our directed acyclic graph can be read as $DAG\ G = \{V, E\}$, where our vertices represent the observed variables $V = \{X_1, ..., X_n\}$, our edges represent the ordered pairs of our nodes $E = \{(A, B)\}\ where\ A, B\ \epsilon\ V,\ and\ a \neq b$.
Each entry of the probability table of each node could be quantified. As marginal probability for root nodes $P(X = x_i)$ ,and as conditional probability for non-root nodes

$$P(X_1 = x1, ..., X_n = x_n) = \prod_{i=1}^{n} P(x_i | Parents(X_i))$$

But to achieve this ability to quantify conditional probabilities, with such low complexity operations compared to the operations on a full joint distribution, the structure of the graph have to be inferred.
The task of learning the structure of such a Bayesian network from a dataset, could be done through different approaches, local and global score based learning, and the mixtures of those two.However, to enable such learning, we need to make certain assumptions.

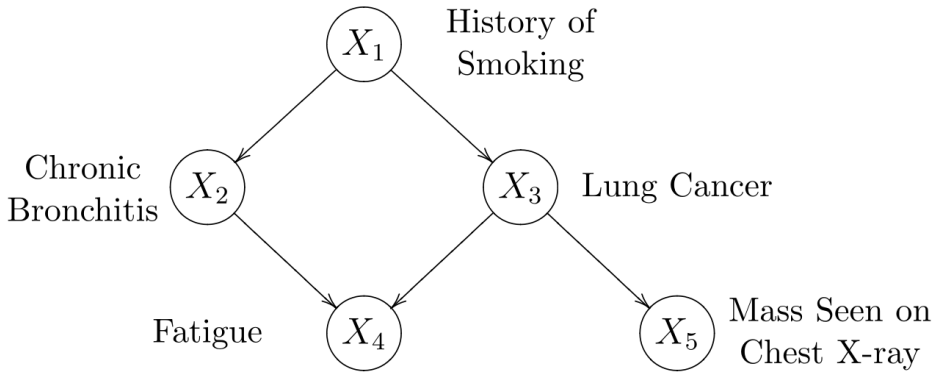



Figure 2: A hypothetical causal Bayesian network structure, Source: Mani 2006

## 5.2 Assumptions

Certain algorithms, score functions require further assumptions which will be defined along the definition of those functions. In this section the base assumptions are listed, with the corresponding theorems that are required before the modelling phase itself.

**Theorem 5.1 ( Ordered Markov Condition )**
*A necessary and sufficient condition for a probability distribution $P$ to be Markov relative a $DAG\,G$ is that, conditional on its parents in $G$, each variable $X_i$ must be independent of all its predecessors in some ordering of the variables that agrees with the directed edges of $G$. (Pearl 2010)*

**Theorem 5.2 ( Faithfulness Condition )**
*Suppose we have a joint probability distribution P of the random variables in some set $V$ and a $DAG\,G = (V,\, E)$. Then $(G,\, P)$ satisfies the faithfulness condition if and only if all and only conditional independences in $P$ are identified by d-separation (as seen in definition 5.2.1) in $G$. (Neapolitan 2003)*

### 5.2.1 Causal Markov Assumption

Using the theorem 5.1, if we create a causal graph $DAG\,G = \{V, E\}$ and assume that the probability distribution $P$ for each variable $X_i$ in $V$ to satisfies the Markov Condition, that translates into that we are making the Causal Markov Assumption.
The assumption itself can be understood as each variable $X_i$ in $V$ is independent of all its non-descendants, given its $parents(X_i)$ in $G$.

### 5.2.2 Causal Faithfulness and Stability Assumption

Opposed to the Causal Markov Condition which specifies the independence relationships among each variable $X_i$ in $V$, the Casual Faithfulness Condition focuses on the dependence

relationships to specify them. Generally, we would want each edge in $E$ to mean that there is a direct dependency.
Essentially we assume that variables $V$ are dependent unless their independence is implied by the Causal Markov Condition as seen in theorem 5.1.

### 5.2.3 Statistical Testing Condition

A statistical test performed to determine dependence given a finite dataset will be correct relative to the dependence in the joint probability distribution, that is defined by the causal process under study. Since our datasets contain only a finite set of records, the relationships among the variables cannot be known with absolute certainty. Therefore an assumption has to be made, that the dependence or independence relationship that has been inferred is the same, to the relationship that would have been inferred, in a case of infinite records. (Fehér 2018) Let $P(X)$ be the probability distribution we would have inferred, having the sample size of our variable at infinity $|n| = \infty$, as our sample size increases, our probability distribution $P(X')$ approaches $P(X)$, meaning :

$$\lim_{|n|\to\infty} P(X') = P(X)$$

## 5.3 Structure learning

Learning Bayesian networks are not easy, it has been shown that finding a general Bayesian network, or even just finding an approximate solution is an NP-hard problem Carvalho and Ados 2019. Nonetheless the achievement of acceptable result, is approachable from two directions.
Either considering the whole structure of the network while directing edges, or trying to propagate them, in a local manner, focusing on the nature of a subset of variables. The latter is called local structure learning, while the other is global learning.

### 5.3.1 Local Causal Discovery

In Local Causal Discovery the main goal is to infer local sub-models, i.e. inferring pairwise causal directions on our observed variables only using a subset of all variables.
In its simplest form three variables are enough to infer such a direction. Although it is necessary to mention that such local algorithms are less complex, and thus have a much shorter runtime, they do not strive to infer the full causal structure, and thus they might not reproduce the true causal structure accurately.
The LCD algorithm (Mani and Cooper 2004) is based on the attributes of variable triplets, inferred through d-separation (presented in definition 5.2.1 below). Basically it utilizes statistical dependence and independence testing of variables, and allows to infer some casual structures.
Such algorithms could test for two different types of structures, V-structures (described in section 5.3.2) and LCD structures (described in section 5.3.3). A considerable issue

with LCD structures (chains) is that due to observational equivalence they cannot be distinguished from one another. This results in undirected edges in the model, although V-structures could be inferred with reasonable certainty. Such undirected edges made from LCD structures, could be directed by using the edges of V-structures, and applying constraints.
Among those constraints, one basic attribute of causal graphs is always present. There can be no cycles. Other constraints could be present in the from of expert knowledge, e.g. already known ground truths on the relationships of the variables.

**Definition 5.2.1 ( d-separation )**
*A path $p$ is said to be d-separated (or blocked) by a set of nodes $Z$ if and only if*

- *$p$ contains a chain $i \leftarrow m \leftarrow j$ or a fork $i \rightarrow m \leftarrow j$ such that the middle node $m$ is in $Z$ or*
- *$p$ contains an V-structure (or inverted fork) $i \leftarrow m \rightarrow j$ such that the middle node $m$ is not in $Z$ and such that no descendant of $m$ is in $Z$.*

*A set $Z$ is said to d-separate $X$ from $Y$ if and only if $Z$ blocks every path from a node in $X$ to a node in $Y$. (Pearl 2010)*

### 5.3.2 V-structure

Given three variables if and only if the following conditions are met, then the underlying causal structure is considered as a V-structure ($X \rightarrow Y \leftarrow Z$) seen on Figure 3. Where $Y \not\perp\!\!\!\perp X$ means that $Y$ and $X$ are not independent variables. (Mani and Cooper 2004)

$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp Z$$
$$X \perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Z|Y$$

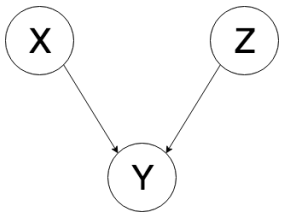
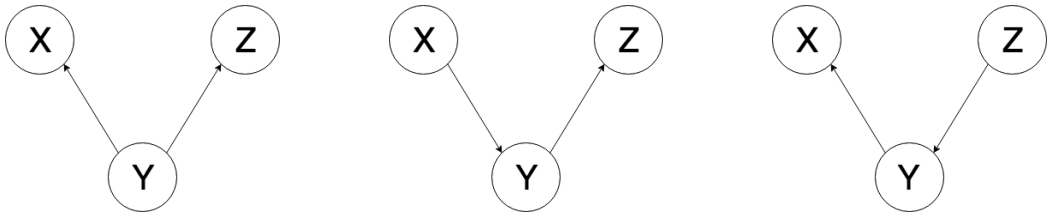



Figure 3: V-structure

### 5.3.3 LCD structure

Similarly, we could infer LCD structures as well, although in this case, we could only acquire the information that there exists a relationship, among variable $X, Y, Z$ but on the direction of the edges, we could gain no further information. The edges might form three different types of structure as seen on Figure 4. One of the following three structures $\{(X \leftarrow Y \leftarrow Z),\ (X \rightarrow Y \rightarrow Z),\ (X \leftarrow Y \rightarrow Z)\}$, could be an underlying structure if and only if the following conditions are considered true. (Mani and Cooper 2004)

$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp Z$$
$$X \not\perp\!\!\!\perp Z$$
$$X \perp\!\!\!\perp Z|Y$$




(a) Fork (b) Chain (c) Reverse chain

Figure 4: LCD structures

### 5.3.4 Dependence Testing

Such dependence testing that is required in the methods described in section 5.3.2 and 5.3.3, can be done through the means of statistical hypothesis testing. Without being exhaustive, dependence testing for discrete variables could be done through the means Chi Square Score $(\chi^2)$ (McHugh 2013).It is worthy to mention, that Chi Square comes with it's own set of assumptions, that are not detailed here.
The score itself could be calculated through the following (Fehér 2018):

$$\chi_c^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where

- $c$ means the degree of freedom, essentially the multiplication of the two variables cardinality of the set of values, that is being tested, minus one $(n-1) * (m-1) - 1$.
- $O$ means the observed value. The count of cases in each cell of the table.
- $n$, $m$ the cardinality of the set of values of the two tested variables.

- $\chi^2$ the computed $\chi^2$ value.
- $E$ is the expected value, calculation as shown below.

$$E = \frac{M_R * M_C}{|n|}$$

Where

- $M_R$ represents the row marginal for that cell,
- $M_C$ is the column marginal for that cell,
- $|n|$ is the normalizing constant, in our case the number of rows in our dataset.

Using this formula allows the computation of chi square $\chi^2$ value which can be compared to a critical value defined by the selected significance level and $c$ degree of freedom. Thus this method can be used to test our hypothesis. For example, it can be used to test whether $W$ is not independent of $X$ (in section 5.3.2). That probability of observing the sample statistic as extreme as the test statistic is called the $P$-value of statistics.

## 5.4 Score based learning

The other dominant approach in causal structure inference, is the score based one. In this field, there always exists a score function, that is able to describe the goodness of a given Bayesian network, based on its ability to describe a given dataset.
There are numerous score functions defined, and there are even more techniques that are trying to do optimization using the functions themselves. Without being exhaustive the following sections will include a few approaches, and a few score functions. These will include only information theory based metrics, while Bayesian metrics will be left out, because they are outperformed in several scenarios (Carvalho and Ados 2019).

### 5.4.1 Score functions

Information theory based metrics are based on data compression. Scores falling into this category on average are based on two components, a score part and a penalty part.

$$\phi(G, T) = Score(G, T) + Penalty(G, T)$$

### 5.4.2 Log-Likelihood

The Log-Likelihood ($LL$) is often used as a scoring function, which quantifies the probability of our dataset ($T$), given our Bayesian network ($G$) ,i.e. that the dataset $T$ was generated using the Bayesian network $G$. Also to acquire a more manageable score, a natural logarithm is applied on it's value. (Carvalho and Ados 2019)

$$Score(G, T) = LL(G, T) = log(P(T|G)) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} log(\frac{N_{ijk}}{N_{ij}})$$

where

- $N_{ijk}$ is the number of instances in the data $T$ where the variable $Xi$ takes it's $k$-th value $x_{ik}$ and the variables in the set of $parents(X_i)$ take their $j$-th configuration $w_{ij}$.
- $N_{ij}$ is the number of the instances in the data $T$ where the variables in $parents(X_i)$ take their $j$-th configuration,
$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$
- $w_{ij}$ $(1 <= j <= q_i)$ is a possible configuration of the parents set $parents(X_i)$ of the random variable $X_i$, that has at most $qi$ number of possible configurations.

By itself this score function is usually not used, since it lacks any penalty for complex structures, so it tends to produce not representative complex models.

### 5.4.3 Akaike information criterion

One of the simplest penalties we could introduce is a metric based on the complexity of our network.
Such correction would define a penalty based on the number of parameters $|N|$ that would be required to describe $G$ graph. Which would result in:

$$\phi(G,T) = LL(G,T) - |N|$$

Where the number of parameters $|N|$ is to be understood as

$$|N| = \sum_{i=1}^{n}(s_i - 1)q_i$$

Where $s_i$ means, the finite number of states in $X_i$ and $q_i$ means the number of possible configurations of the parent set $Parents(X_i)$. (Carvalho and Ados 2019)
Essentially meaning, if a node in graph $G$, has numerous parent nodes, it would need a more complex description, that would result in a bigger penalty. (Fehér 2018)

### 5.4.4 Corrected Akaike information criterion

Researchers have shown that AIC on small sample sizes, does not correctly perform the penalization for big parameter size. (Velsen 2009)
But with a correction score, this could be eased, resulting in a score that is almost identical to the AIC score:

$$\phi(G,T) = LL(G,T) - |N| - \frac{2|N|^2 + 2|N|}{|n| - |N| - 1}$$

Where it can be seen that with greater sample sizes, the correction part converges to zero:

$$\lim_{|n|\to\infty} \frac{2|N|^2 + 2|N|}{|n| - |N| - 1} = 0$$

### 5.4.5 Schwarz's Bic and MDL

Based on the principle that the simplest model most likely be the right one, there is a score function defined. Where the penalty part is almost similar to the AIC score function's penalty part, although it has been multiplied by the log of the sample size, which essentially means the number of bits needed to encode the parameters. In our special case the two metrics, Schwarz's Bayesian information criterion and Minimum Description Length do not differ. (Nir Friedman 1999)

$$\phi(G, T) = LL(G, T) - \frac{1}{2} * log(|n|) * |N|$$

## 5.5 Optimization

Using score functions by themselves with exhaustive search to learn the underlying causal structure is not the optimal way to proceed. Instead, combining score functions with optimization techniques could yield acceptable results. This can be formulated as follows:

**Definition 5.2.2 ( Learning a Bayesian network )**
*Let $B_N$ be the set of Bayesian networks with $N$ variables.*
*Given a Network $G$, a data $T = y_1, ..., y_N$ and a scoring function $\phi$, the problem of learning a Bayesian network is to find a Bayesian network $G \in B_N$, that maximizes the value $\phi(G, T)$. (Carvalho and Ados 2019)*

For example, in algorithm 1 a simple greedy Hill climbing algorithm is used as an optimization technique with Bayesian network structure learning. (Russell and Norvig 2009)

**Data:** $G = \{V, E\},\ T$
**Result:** $G' = \{V, E'\}$
score_increasable = $True$;
edges = $E$;
vertices = $V$;
**while** *score_increasable* **do**
    Score = $0$;
    **for** *i in all_possible_edge_changes*$(V, E)$ **do**
        $E' = change(E, i)\}$;
        **if** $\phi(G' = \{V, E'\}, T)$ > *Score* **then**
            Score = $\phi(G' = \{V, E\}, T)$;
        **end**
    **end**
    **if** *Score* $\neq 0$ **then**
        $E = E'$;
    **else**
        score_increasable = $False$;
    **end**
**end**
return $G = \{V, E\}$;

**Algorithm 1:** Hill climbing algorithm on Bayesian networks

It is worthy to mention that greedy algorithms are more prone to stuck in a local optimum. In other words, they tend not to find the optimal score in some cases. Fortunately, there are multiple ways to remedy that. At first we could infer a causal skeleton, i.e. run a local dependence search, and find all LCD and V-structures, and then optimize the structure from that starting point (Tsamardinos, Brown, and Aliferis 2006), or we could include expert knowledge.
Although the replacement of the optimization technique is possible as well, since there exists multiple methods that could be used, for example gradient based searches or genetic algorithms (Russell and Norvig 2009).

## 5.6 Receiver operating characteristic

Measuring the goodness of causal graph learning algorithms can be performed in multiple ways. A possible option is to use ROC AUC metrics. The goodness can be represented with a graphical plot that is able to illustrate the abilities of a classifier, in our case the abilities of our algorithm to classify the direction of an edge. The calculations are rather straightforward although the variables need to be defined, such as as below. True positive ($TP$) is the count of the edges that are in the underlying causal graph, and in the inferred graph as well. While its opposite is the False positive ($FP$) edges count, which is the number of edges that are not in the original structure, but were detected as edges by the algorithm.

True negative ($TN$) sums the amount of edges that are equally not present in both of the graphs, while False negative ($FN$) sums the ones that are in the original causal graph, but were not detected by the algorithm. (Heijden, Velikova, and Lucas 2013)

| Sum | Condition Positive ($P$) | Condition Negative ($N$) |
|---|---|---|
| Prediction Positive ($P$) | TP: Edges present in both graphs. | FP: Edges wrongly taken. |
| Prediction Negative ($N$) | FN: Edges wrongly missing. | TN: Edges not present in both graphs. |

Table 1: Confusion matrix

Based on these four measures different metrics could be derived. (Russell and Norvig 2009)

| |
|---|
| $Positive(P) = TP + FN$ |
| $Negative(N) = TN + FP$ |
| $Accuracy = \frac{TP + TN}{P + N}$ |
| $True\ Positive\ Rate\ (TPR) = \frac{TP}{P}$ |
| $False\ Positive\ Rate\ (FPR) = \frac{FP}{N}$ |
| $Area\ Under\ the\ Curve\ (AUC) = \sum_{i \in FPR} TPR_i$ |

Table 2: Derivations from the confusion matrix

Where semantically Positive means the edges that are present in the original graph, and Negative means the opposite, i.e. those edges are not present. Accuracy could be understood as how well the classifier could classify the edges. On one hand, True Positive Rate focuses only on how well we could classify the edges that are originally present, while the opposite False Positive Rate tells us how well we the algorithm noticed the edges that should not be taken.

## 5.7 Distributions

To test a function, that would be able to propagate the causal order between variables, on synthetic data, the tests itself somehow have to emulate the diversity of real world observed stochastic variables. In order to approximate that, a few probability distributions were investigated, namely the ones listed in the subsection. To gain better insight their mass functions are listed, with a few details. The list was written with the help of the descriptions presented in R Core Team 2019 and in Ketskeméty 1996.

### 5.7.1 Uniform distribution

The uniform distribution occurs when our random variable $X$ on an interval $]a; b[$ where $-\infty < a < b < \infty$ has a density function of

$$P(X) = \frac{1}{b-a}$$

Where the mean is $\frac{1}{2}(a+b)$ and the variance is $\frac{1}{12}(b-a)^2$.

### 5.7.2 Poisson distribution

The Poisson distribution has a mass function of

$$for\ X = \{0, 1, 2, ...\}\ and\ \lambda > 0,$$

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where the mean and the variance are equally $\lambda$.

### 5.7.3 Geometric distribution

The Geometric distribution for $X = \{0, 1, 2, ...\}$ and $0 < p <= 1$ where the density is

$$P(X = k) = p(1-p)^k$$

With a mean of $\frac{1-p}{p}$ and variance of $\frac{1-p}{p^2}$.

### 5.7.4 Multinomial distribution

The Multinomial distribution where $x$ is a $K$ component vector, and $p_1 ... p_k$ are event probabilities has a mass function of

$$P(X_1 = x_1, ..., XK = xk) = \frac{(\sum x_i)!}{x_1! ... x_k!} p_1^{x_1} ... p_k^{x_k}$$

Where the mean is $np_i$, and the variance is $np_i(1-p_i)$.

### 5.7.5 Negative binomial distribution

The Negative Binomial distribution with $\Gamma$ as the gamma distribution and $X = 0, 1, 2, ...$ and $n > 0$ and $1 >= p > 0$ has a mass function of

$$\frac{\Gamma(x+n)}{\Gamma(n)x!}p^n(1-p)^n$$

Where the mean is $\frac{pr}{1-p}$ and the variance is $\frac{pr}{(1-p)^2}$.

### 5.7.6 Hypergeometric distribution

The hypergeometric distribution is used for sampling without replacement. The distribution itself is defined as for $X = \{0, ..., k\}$ with a density function as

$$P(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

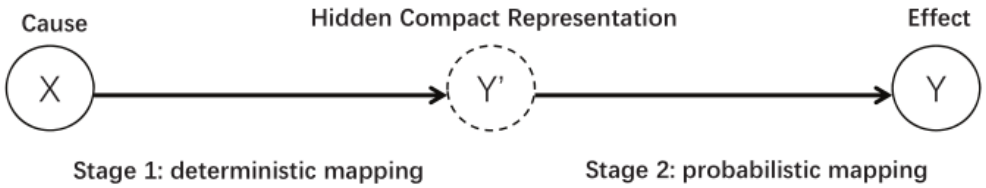
Where the mean is $kp$ and variance is $kp(1-p)\frac{m+n-k}{m+n-1}$.

# 6 Methods and Testing

The testing included the comparison of synthetic datasets, and real world datasets, with methods used in Cai et al. 2018. The goal was to compare the performance of different score functions shown in section 5.4.1 replacing the HCR algorithm's originally used score function described below.
All of the testing, was performed in the statistical language $R$ using the following three packages **HCR, data.table, bnlearn**. **HCR** is the cran package provided by Cai et al. 2018, while **data.table** is a package used for enhanced data storage in tables, and **bnlearn** which is generic Bayesian network package, with all sorts of features.
In the following sections, the accuracy of various methods is measured using **bnlearn**, and inference is facilitated through **hcr**.

## 6.1 Hidden Compact Representations

The method of hidden compact representations, builds on both the world of constraint based methods and the score based ones. The method itself aims to provide a framework of causal discovery on discrete, especially categorical data.
Inference is performed through a two stage process as shown in Figure 5, using the HCR model which is $M : X \rightarrow Y' \rightarrow Y$ where $M$ means our model, $X$ and $Y$ are present variables assumed to be cause and effect, and $Y'$ is a hidden representation. The first step is mapping the cause $X$ to a lower cardinality hidden representation $Y'$, encoding key information, while the second step tries to determine the effect. Through encoding the relevant information to a hidden variable, leaving out irrelevant data, the cause and effect gains a compact representation.




**Stage 1: deterministic mapping**
from cause (X) to hidden representation (Y')

| X \ Y' | Poisonous | Not Poisonous |
|---|---|---|
| Poisonous Mushroom | 1 | 0 |
| Mushroom | 0 | 1 |
| Rice | 0 | 1 |
| Poisonous Fish | 1 | 0 |

**Stage 2: probabilistic mapping**
from hidden representation (Y') to effect (Y)

| Y' \ Y | Food Poisoning | Stomach Flu | Normal |
|---|---|---|---|
| Poisonous | 0.85 | 0.10 | 0.05 |
| Not Poisonous | 0.03 | 0.07 | 0.90 |

Figure 5: Food Poisoning: A Hidden Compact Representation Example in Real World. Source: Cai et al. 2018

The model by itself has certain assumptions about the data itself, namely, that the cause $X$ can be reduced to a hidden low-cardinality space $Y'$, or as seen in the paper itself as assumption A1 which states that
"There does not exist values $y_1 \neq y_2$ such that $P(Y = y_1|X) = P(Y = y_2|X) * a$ for all possible $X$ values. (Note that both $P(Y = y_1|X)$ and $P(Y = y_2|X)$ are functions of X.)"
though even with working with these assumptions could yield acceptable results.

### 6.1.1 HCR model

First, there has to be an understanding of the model itself, to be able to build upon it.
Let $X$ and be the cause of $Y$ in a discrete cause-effect pair $X \to Y$, Where a model $M : X \to Y' \to Y$ could be used to describe the causal mechanism behind the provided data, with $Y'$ as a hidden compact representation. $T = \{(x_i, y_i)\}_{i=1}^{|n|}$ as our data, or group of observations. The log-likelihood of such a model could be described as seen in section 5.4.2, although in this case, it is more descriptive to use the form that is used in Cai et al. 2018, which is the following:

$$\phi(M, T) = log \prod_{i=1}^{|n|} \sum_{y'_i} P(X = x_i, Y' = y'_i, Y = y_i|M)$$

As seen in section 5.3.2, according to $X \perp\!\!\!\perp Y|Y'$, the joint probability could be decomposed into three equations.

$$\phi(M, T) = log \prod_{i=1}^{|n|} \sum_{y'_i} P(X = x_i)P(Y' = y'_i|X = x_i)P(Y = y_i|Y' = y'_i)$$

The first step in the two stage process starts with variable $X$ mapped to a low-cardinality hidden variable $Y'$, using $Y' = f(X)$ where $f : Z \to Z$ is a noise-free arbitrary function. Note that this comes with the assumption that variable $X$ can be reduced to a lower cardinality space $Y'$. Which would not always be true. Although building upon this assumption,the $P(Y' = y'_i|X = x_i)$ part of the equation denotes how the compact representation is generated. Since it is a deterministic process, $P(Y' = y_i|X = xi) = 1$ if $y_i = f(x_i)$ and $P(Y' = y_i|X = x_i) = 0$ if $y_i \neq f(x_i)$ where $f(x)$ denotes a true mapping function. Thus after that, our function could be written up as

$$\phi(M, T) = log \prod_{i=1}^{|n|} \sum_{y'_i} P(X = x_i)P(Y = y_i|Y' = f(x_i))$$

Since this framework, acts the same as it is described in section 5.4.2, and contains a variable with unknown cardinality. It is wise to introduce a penalty part to control the complexity of our model. For instance using the formula described in 5.4.5 an additional penalty part would result in

$$Penalty(M, T) = \frac{(|X| - 1) + |Y'|(|Y| - 1)}{2} * log(|n|)$$

meaning Score would be

$$Score(M,T) = log \prod_{i=1}^{|n|} \sum_{y_i'} P(X = x_i) P(Y = y_i | Y' = f(x_i))$$

and our scoring function as the sum of the two above.

$$\phi(M,T) = Score(M,T) + Penalty(M,T)$$

The model with the highest score $\phi(M,T)$ is considered the best candidate to recover the causal model. Where an alternating maximization procedure was used to gain the list of score pairs of $T = (\{x_i, y_i'\})_{i=1}^{|n|}$, and thus to infer $Y'$. Although this recovery is not the focus of this paper.
The direction of the edge between $X$ and $Y$ could be inferred based on the calculation of the score given on two different models, as $M : X \to Y' \to Y$ and $M' : Y \to X' \to X$, where

- $\phi(M,T) > \phi(M',T)$, infer $X \to Y$.
- $\phi(M,T) < \phi(M',T)$, infer $Y \to X$
- $\phi(M,T) = \phi(M',T)$, infer non-identifiable.

In the following sections, testing is performed by replacing of the above given $Penalty(M,T)$ score with the below listed penalty functions.

| |
|---|
| Log: $Penalty(M,T) = 0$ |
| Aic: $Penalty(M,T) = \|N\|$ |
| Aicc: $Penalty(M,T) = \|N\| - \frac{2\|N\|^2 + 2\|N\|}{\|n\| - \|N\| - 1}$ |
| BIC: $Penalty(M,T) = \frac{1}{2} * log(\|n\|) * \|N\|$ |

Table 3: Penalty functions

## 6.2 Simulated data

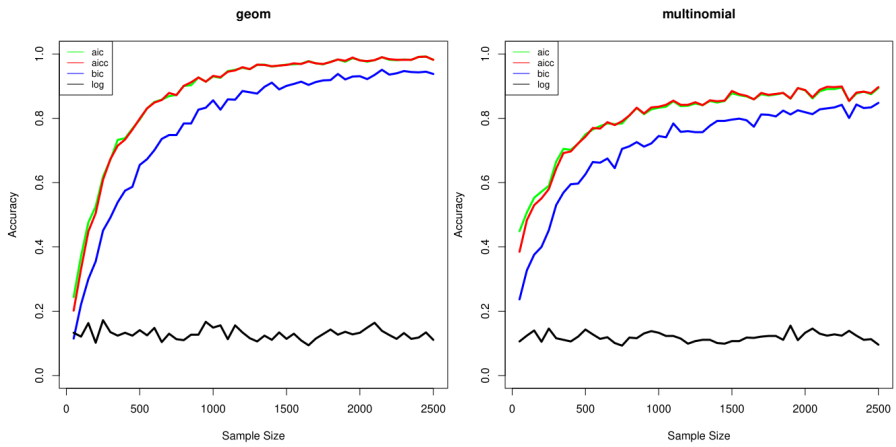
Similar as seen in Cai et al. 2018, the sensitivity to sample size was tested by the accuracy measured against the sample size. The Synthetic data was generated in the same manner as seen in the original paper, using the **HCR** cran package.

For each distribution stated in section 5.7 the following was done for different sample sizes $|n|$. At first variable $X$ from the given distribution, with a randomly chosen cardinality from $3, ..., 15$ was generated with the sample size $|n|$. Next, each sample from $X$ was mapped to a value that uniformly samples from the interval $\{1, 2, ..., |X|\}$. Finally, a conditional probability distribution $P(Y|Y')$ was generated randomly, and $Y$ was sampled accordingly to $Y'$ and $P(Y|Y')$. Also $|Y|$ was generated from the interval $\{|Y'|, ..., 15\}$.
Then the different score functions of section 5.4.1 were evaluated as the HCR algorithm's scoring function on the generated dataset, both on $X \to Y$ and $Y \to X$ direction. This process was repeated $1000$ times then the results were averaged as seen in the following algorithm 2.
It is worth mentioning, that the algorithm is compactly represented to make it easier to read,but it was ensured that each score function would run on the same dataset at each iteration.

**Data:** SampleSize, $\phi$, Distribution
**Result:** Average
i = 0;
Average = 0;
**while** *i < 1000* **do**
  dataset = GenerateDataset(SampleSize,Distribution);
  ScoreXY = $\phi$(dataset$X, dataset$Y);
  ScoreYX = $\phi$(dataset$Y, dataset$X);
  **if** *ScoreXY > ScoreYX* **then**
    Average += 1;
  **end**
**end**
return Average/1000;

**Algorithm 2:** Evaluation of Score functions on generated data on a given distribution

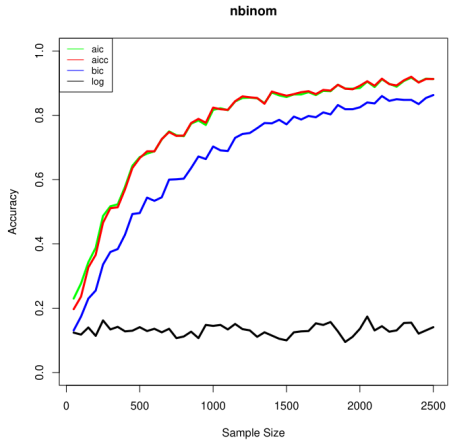
The accuracy scores for different variable distributions can be seen on the following figures. Results indicate that compared to the original score presented in the paper, (namely BIC), AIC and AICC provide an increased percentage of accuracy in case of every distribution.
This synthetic testing is a corner case, which essentially means that the algorithm has been given nodes $A$ and $B$, with a certain underlying causal structure, and it has to decide it's direction. The result is either $A \to B$, $B \to A$, or non-identifiable.
The paper did not state how the algorithm performs when there is no causal relationship between the variables. Such examples can be seen on figure 7, which will be useful on interpreting real world datasets.



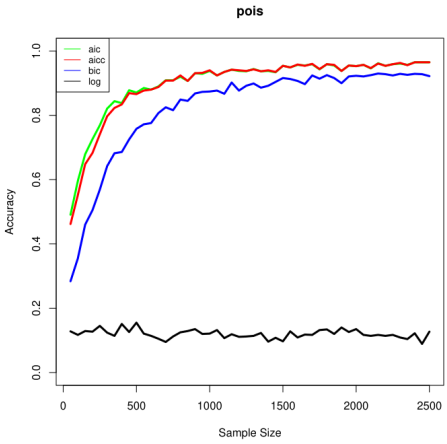

(a) Geometric distribution



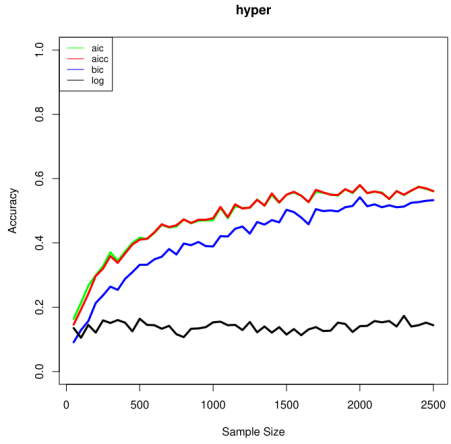

(b) Multinomial distribution



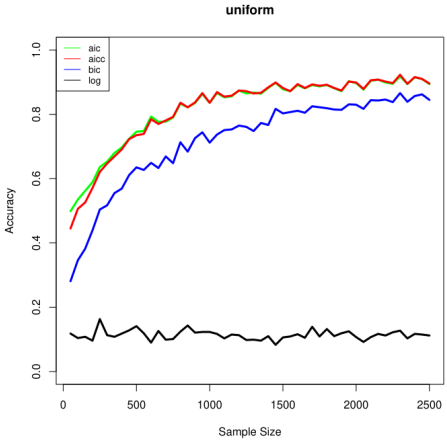

(c) Negative binomial distribution




(d) Poisson distribution




(e) Hypergeometric distribution




(f) Uniform distribution

Figure 6: Accuracy of different score functions given different variable distributions.

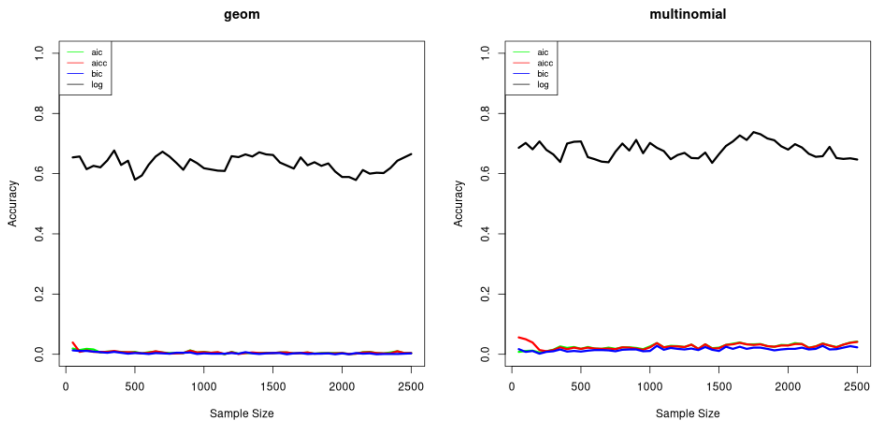
## 6.3 Simulated Noise

The performance testing was also conducted on noise, where the algorithm tries to infer causal relationship between $X$ and $Y$, in the manner as shown in 3. In this case $X$ and $Y$ will be random data separately generated from the given distribution. Results show that the HCR algorithm, which is able to direct an edge when there surely is a connection between two variables, performs terribly when the task is the detection of the absence of a causal relationship.

**Data:** SampleSize, $\phi$, Distribution
**Result:** Average
i = 0;
Average = 0;
**while** *i < 1000* **do**
    dataset = GenerateDataset(SampleSize,Distribution);
    ScoreXY = $\phi$(dataset$X, dataset$Y);
    ScoreYX = $\phi$(dataset$Y, dataset$X);
    **if** *ScoreXY == ScoreYX* **then**
        Average += 1;
    **end**
**end**
return Average/1000;

**Algorithm 3:** Evaluation of Score functions on generated noise on a given distribution

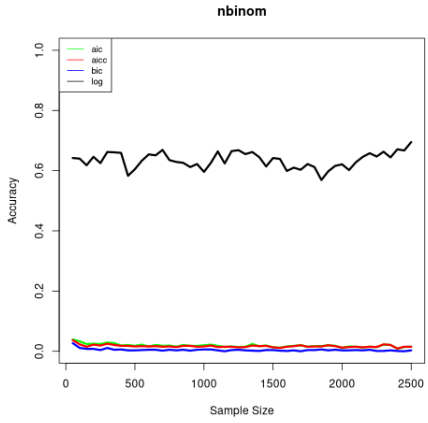
The results shown in 7 might be explained by its inherent property, i.e. it orients edges towards the higher score. However the probability that scores corresponding to the two directions would end up on the same value is relatively low. Thus the probability of a non-defined result is low. For future research it is worthy to note the introduction of a minimum $\delta$ distance between the two scores could mitigate the problem considerably.



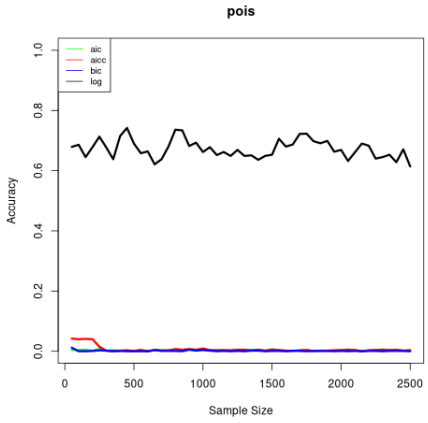

(a) Geometric distribution



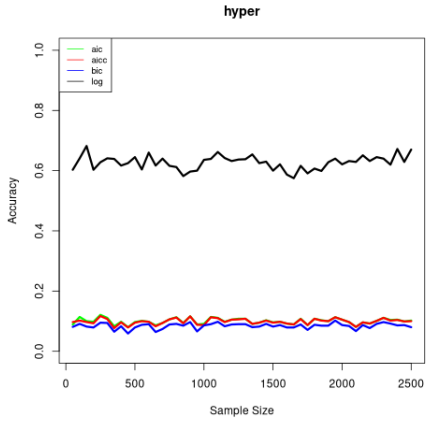

(b) Multinomial distribution



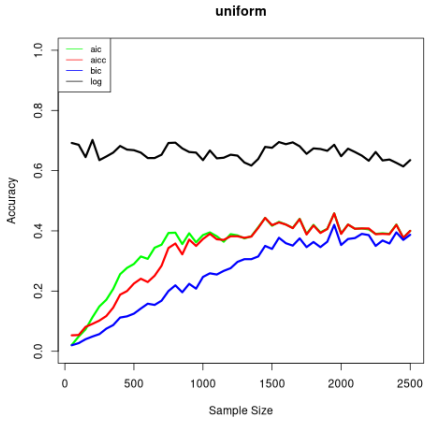

(c) Negative binomial distribution




(d) Poisson distribution




(e) Hypergeometric distribution




(f) Uniform distribution

Figure 7: Accuracy of different score functions on noise of different distributions.

## 6.4 Real-world dataset

The HCR algorithm was tested on real world datasets as well, although with mixed results.

### 6.4.1 Abalone

On the abalone dataset, the algorithm was not able to identify the underlying ground truths. The directions in the original paper resulted were $Sex \to Length$, $Sex \to Diameter$, $Sex \to Height$. However the tests indicated the opposite direction for all three relationships ( $Length \to Sex$, $Diameter \to Sex$, $Height \to Sex$). Even the ones in the original paper lead to poor results. The only exception was the simple loglikelihood, which was not able to detect the direction at all.

### 6.4.2 Bridges

The Bridges dataset consists of 13 variables and 4 causal relationships. Using the same metric as for the analysis of the synthetic data sets, accuracies were measured for various sample sizes (see Figure 8). Interestingly, in contrast with the analysis of the synthetic data set, the loglikelihood score performed the best. That can be explained examining the other metrics.

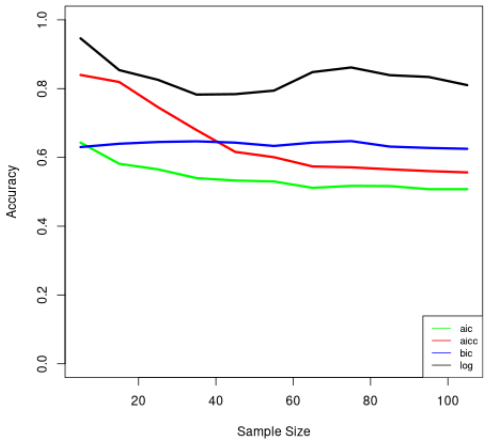
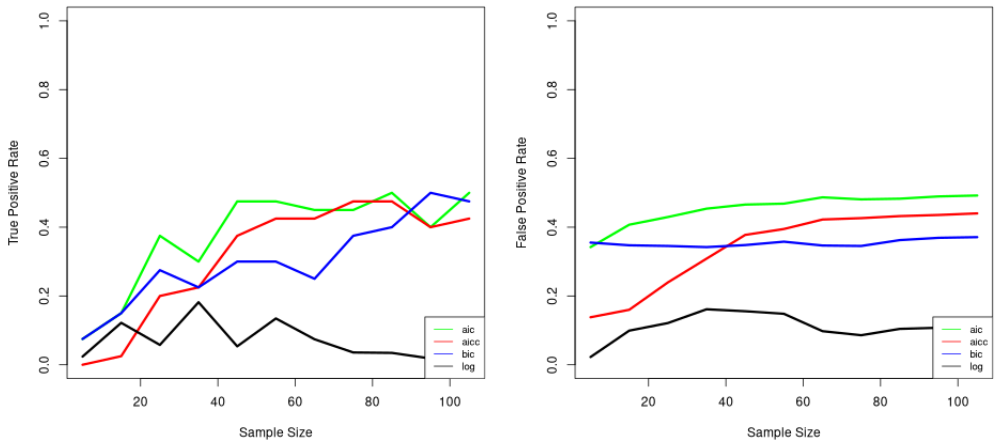



Figure 8: Accuracy on Bridges

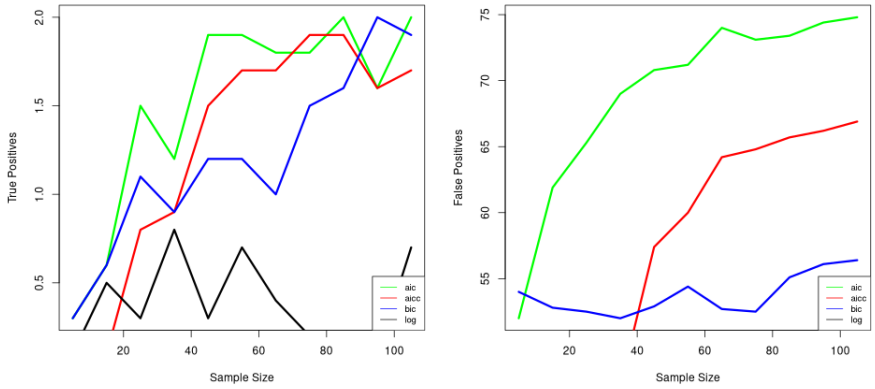
It is the true that the rate of true positives increase, but at the same time the rate of false negatives stays the same, or increases slightly. The reason behind this could be better seen from the graph of true positives and false positives (See Figure 9).




(a) True Positive Rate (b) TPR, FPR on Bridges

Figure 9

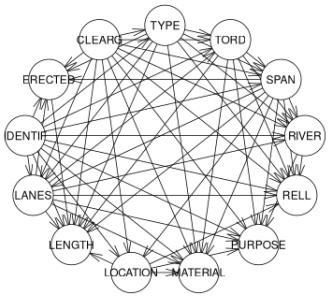
Here on Figure 10 it can be seen that, the algorithm's ability to successfully direct the edges increases with larger sample sizes, but it also increases its rate of finding false positive edges Note that in each graph in almost every case the aic scoring leads.



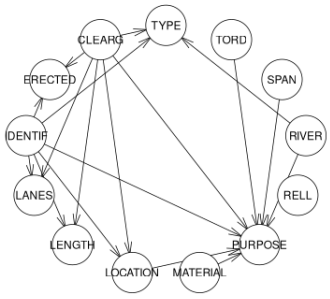

(a) True Positives (b) False Positives

Figure 10: TP, FP on Bridges

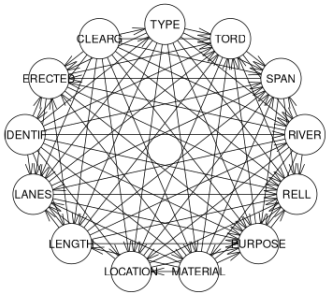
In contrast, and for better visualization, the inferred graphs can be seen on Figure 11. Also there is the graph included with the ground truths, that can be seen on figure 12.



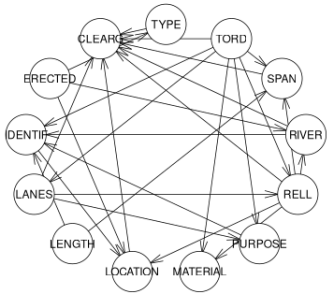

(a) Aic on 10 samples




(b) Log on 10 samples




(c) Aic on 108 samples




(d) Log on 108 samples

Figure 11: Graph inference

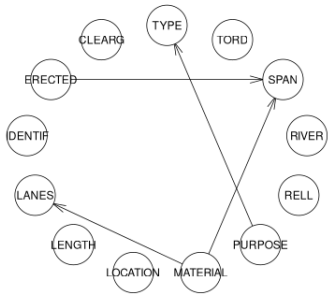



Figure 12: Bayesian network representing the ground truths of the Bridges dataset

# 7 Conclusion

This thesis provided an overview of the necessary theoretical background of Bayesian networks, its structure learning methods, and it's measurement tools. With these the testing and measuring of the properties of the Hidden Compact representations algorithm could be conducted.
The HCR algorithm itself is a powerful tool, and it outperforms many existing algorithms in directing the edges of Bayesian networks, although, it has it's assumptions, and performs exceptionally poorly on cases where edges actually do not exist.
It was originally proposed with a Schwarz's Information Criterion as it's built in scoring function, although after thorough testing it could be seen that in almost every case the Akaike information criterion outperforms it.
In summary, the algorithm itself with Akaike information criterion as it's score function, performs exceptionally good on categorical data in cases where there surely is a causal direction. Therefore, by itself its usage could be questioned, but combined with other algorithms, it can very well increase their edge directional abilities, between discrete variables.

## 7.1 Acknowledgements

I would like to thank the support of Dr. Gábor Hullám whose research has been supported by the ÚNKP-19-4 New National Excellence Program of the Ministry for Innovation and Technology (ÚNKP-19-4-BME-344),

# References


András, Millinghoffer, Gábor Hullám, and Péter Antal (Oct. 2019). “Statisztikai adat- és szövegelemzés Bayes-hálókkal: a valószínûségektôl a függetlenségi és oksági viszonyokig”. In:

Cai, Ruichu et al. (2018). “Causal Discovery from Discrete Data using Hidden Compact Representation.” In: *Advances in neural information processing system*.

Carvalho, Alexandra and Avanç Ados (Oct. 2019). “Scoring functions for learning Bayesian networks”. In:

Fehér, Péter (2018). “Kauzális kapcsolatk tanulása kényszeralapú algoritmusokkal”. In: URL: https://diplomaterv.vik.bme.hu/hu/Theses/Kauzalis-kapcsolatok-tanulasa-kenyszeralapu.

Heijden, Maarten van der, Marina Velikova, and Peter J. Lucas (Dec. 2013). “Learning Bayesian networks for clinical time series analysis”. In: *Journal of biomedical informatics* 48. DOI: 10.1016/j.jbi.2013.12.007.

Ketskeméty, László (1996). *Valószínűségszámítás*.

Mani, Subramani (Mar. 2006). “A Bayesian Local Causal Discovery Framework”. In:

Mani, Subramani and Gregory Cooper (Feb. 2004). “Causal discovery using a Bayesian local causal discovery algorithm”. In: *Studies in health technology and informatics* 107, pp. 731–5. DOI: 10.3233/978-1-60750-949-3-731.

McHugh, Mary (June 2013). “The Chi-square test of independence”. In: *Biochemia medica* 23, pp. 143–9. DOI: 10.11613/BM.2013.018.

Neapolitan, Richard (Jan. 2003). *Learning Bayesian Networks*. DOI: 10.1145/1327942.1327961.

Nir Friedman, Lise Getoor (1999). “Efficient Learning using Constrained Sufficient Statistics”. In:

Patel, Hetal and Dharmendra Patel (June 2014). “A Brief survey of Data Mining Techniques Applied to Agricultural Data”. In: *International Journal of Computer Applications* 95, pp. 6–8. DOI: 10.5120/16620-6472.

Pearl, Judea (2010). *Causality: Models, Reasoning and Inference*. Pearson Education, Inc. ISBN: ISBN-13 978-0-13-604259-4, ISBN-10: 0-13-604259-7.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Russell, Stuart J. and Peter Norvig (2009). *Artificial Intelligence A Modern Approach*. Cambridge University Press. ISBN: 978-0521895606.

Tsamardinos, Ioannis, Laura Brown, and Constantin Aliferis (Oct. 2006). “The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm”. In: *Machine Learning* 65, pp. 31–78. DOI: 10.1007/s10994-006-6889-7.

Velsen, J. L. van (2009). “A corrected AIC for the selection of seemingly unrelated regressions models”. In: arXiv: 0906.0708 [stat.ME].