

Department of Computer Science



Submitted in part fulfilment for the degree of MEng.

Learning Causal Networks in Python

James Callan

Supervisor: James Cussens

Contents

Executive Summary	vi
1 Introduction	1
2 Literature Review	2
2.1 Causal Networks	2
2.2 Probability and Independence	3
2.2.1 Basic Probability	3
2.2.2 Distributions	3
2.2.3 Independence	3
2.3 D Separation	4
2.3.1 Definitions	4
2.3.2 D-Separation	4
2.4 Faithfulness	5
2.5 Discriminating paths	5
2.6 Possible D-Separation Set	6
2.7 Tests for Independence	6
2.7.1 Algorithms For Learning Causal Networks	8
2.7.2 Constraint Based Methods	8
2.7.3 Score Based Methods	8
2.8 Computing a skeleton	9
2.9 The PC Algorithm	10
2.10 The FCI algorithm	11
2.11 The RFCI Algorithm	12
3 Design	15
3.1 Methodology	15
3.1.1 Waterfall	15
3.1.2 Agile	16
3.1.3 Spiral	16
3.1.4 Chosen Approach	16
3.2 Sprints	17
3.3 Requirements	18
3.3.1 Epics	18
3.3.2 User Stories	18
3.3.3 Tasks	18
3.4 Architecture	19
3.4.1 Unit Testing	19

Contents

3.5	Documentation	19
3.6	Tools	20
3.6.1	Version Control	20
3.6.2	Python Libraries	21
3.7	Refactoring	21
4	Implementation	23
4.1	Independence Test	23
4.2	Skeleton	24
4.2.1	Variables to be Tested	24
4.2.2	Order Dependence	24
4.2.3	Separation Sets	24
4.3	PC Edge Orientation	25
4.3.1	V-Structures	25
4.4	FCI Algorithm Edge Orientation	26
4.5	Testing	27
4.5.1	Independence Test	27
4.5.2	Skeleton Estimation	27
4.5.3	PC Algorithm	27
4.5.4	FCI Algorithm	28
4.5.5	Final Structure	28
4.5.6	Independence Test	28
4.5.7	Graph Learners	28
4.5.8	Graphs	29
4.6	Usage	30
4.7	Viewing Graphs	30
4.7.1	Adjacency Matrices	31
5	Analysis	32
5.1	Initial Skeleton Generation	32
5.2	PC Algorithm Edge Orientation	32
5.3	FCI Algorithm Final Skeleton	33
5.4	FCI Algorithm Edge Orientation	34
5.5	Run Speed	34
6	Conclusion	35
6.1	Further Work	35
A	Some appendix	36

List of Figures

2.1	A DAG with 3 nodes and 2 edges	2
2.2	A PDAG with 3 nodes and 2 edges	2
2.3	A graph containing a collider	4
2.4	A DAG in which nodes 1 and 4 are D-Separated	5
2.5	A DAG in which nodes 1 and 4 are not D-Separated	5
2.6	$A - X - Y - Z$ is a discriminating path on Y	6
3.1	Numpy Style docstring	20
4.1	A skeleton of a Causal Network	25
4.2	Two potential colliders in a skeleton	26
4.3	A potential orientation of the skeleton in Figure 4.1 containing colliders from Figure 4.2	26
4.4	A PDAG and how it would be represented in the implementation	29
4.5	A PAG and how it would be represented in the implementation	30
5.1	The actual network that generated data in the asia_1000 dataset	33
5.2	The networks estimated by the R (left) and python (right) implementation from the Asia_1000 dataset	33

List of Tables

2.1	Example Contingency Table For Two Variables	7
-----	-------------------------------------------------------	---

Executive Summary

1 Introduction

The aim of this report is to detail the implementation of various algorithms to learn causal networks in python. These algorithms have already been implemented in the statistical programming language R [1]. However, recently python has become more widely used in the field of statistical computation and machine learning [2].

Python offers a number of advantages over R. Python is much more widely used in general than R [3] so an implementation in python would be accessible to a wider array of developers. Python can also be used in a wider variety of settings than R. R is designed for data analysis and data visualisation, whereas python is designed to be general purpose. This makes it easier in python than in R to incorporate data analysis into wider systems such as web applications.

2 Literature Review

2.1 Causal Networks

Causal Networks are graphical models which represent a set of variables, their conditional dependencies, and their causal relationships [4] as a Directed Acyclic Graph (DAG) or a Maximal Ancestral Graph (MAG). The nodes of the Graph represent the variables. The edges of the graph represent causality, the direction of the edge represents the direction of causality with parent nodes causing child nodes [4].

In a DAG all edges have a single direction, whereas in a MAG edges may be directed or bidirectional [5].

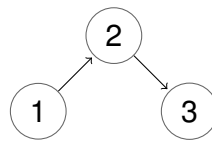


Figure 2.1: A DAG with 3 nodes and 2 edges

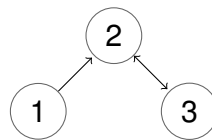


Figure 2.2: A PDAG with 3 nodes and 2 edges

The learning of causal networks allows relationships between variables to be uncovered and presented in a simple and human readable fashion. This can provide useful information for further data analysis as the set of variables that cause another could be used to predict its value. Causal relationships between any pair of variables of a causal network can be easily found.

2.2 Probability and Independence

2.2.1 Basic Probability

In probability theory the confidence that a particular event E occurs can be quantified. This confidence, denoted as $p(E)$ is a real number between 1 and 0. $p(E) = 1$ representing a certainty and $p(E) = 0$ representing E being impossible [6].

In the case of random variables, an event would be a variable X taking a particular value x_i , where x_i is a member of the set $x = \{x_1, x_2, \dots, x_n\}$ of possible values that X can take. The probability of this event is denoted as $p(X = x_i)$ [6].

2.2.2 Distributions

The probability distribution $p(X = x)$ or simply $p(X)$ defines the likelihood of X taking any value. Distributions can be continuous or discrete depending on the variable [6].

Distributions of more than one variable can be described with joint distributions. For variables X and Y the distribution $p(X, Y)$ describes the probability of both X and Y taking particular values simultaneously or $p(X = x, Y = y)$ for every x and every y [6].

Conditional distributions can describe how the value of one variable can effect the value of an other. $p(X|Y)$ describes the probabilities of X taking particular values "given" a value that Y has taken. $p(X|Y)$ is defined as $p(X, Y)/p(Y)$ [6].

2.2.3 Independence

If two variables are independent there is no correlation between the values they take. That is for two variables X and Y , the distribution $p(X = x_i|Y)$ would be the same for all values of Y . Therefore $p(X|Y) = p(X)$. If two variables are independent we can infer that there is no causal relationship between them.

Given the definition $p(X|Y) = p(X, Y)/p(Y)$ and $p(X|Y) = p(X)$ when X and Y are independent, it easy to see that $p(X, Y) = p(X)p(Y)$ if X and Y are independent.

Two variables X and Y can be considered independent conditioned on a third variable Z if there is no correlation between X and Y for given the

value of Z .

In this case $p(X|Y, Z) = p(X|Z)$, as the value of Y has no impact on the value of X . The definition of conditional distributions shows $p(X|Y, Z) = \frac{p(X, Y, Z)}{p(Y, Z)}$, and $p(Y, Z) = p(Y|Z)P(Z)$. Using these definitions and basic algebra $p(X, Y|Z) = p(X|Z)p(Y|Z)$ can be shown when X and Y are conditionally independent on Z .

2.3 D Separation

2.3.1 Definitions

A path is any sequence of adjacent edges regardless of their directionality.

A collider is node in a path which is both entered and left on edges which are directed toward the node.

Unblocked refers to a path that does not traverse a collider [7].

A triple is a set of three nodes on a graph. An unshielded triple is a triple $\langle i, j, k \rangle$ in which there are edges $i - j$ and $k - j$ but no edge $i - k$.

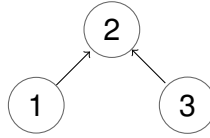


Figure 2.3: A graph containing a collider

2.3.2 D-Separation

If every path between nodes X and Y traverses a collider, nodes X and Y are unconditionally d-separated or d-separated conditioned on the empty set [7]. Two unconditionally d-separated nodes in a causal network are considered to be independent [8].

Two nodes X and Y are d-separated conditioned on set Z if:

1. There exists no unblocked path from X to Y that does not traverse any members of Z . [7]
2. There is no blocked path from X to Y in which all colliders are members of Z or have descendants in Z . [7]

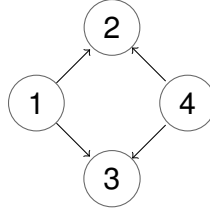


Figure 2.4: A DAG in which nodes 1 and 4 are D-Separated

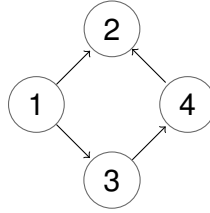


Figure 2.5: A DAG in which nodes 1 and 4 are not D-Separated

Causal networks represent conditional independence with d-separation, variables X and Y are conditionally independent on set Z if nodes X and Y are d-separated conditioned on nodes in set Z [4].

2.4 Faithfulness

By testing for d-separation of variables X and Y given conditioning set Z on a graph we can see if $X \perp\!\!\!\perp Y|Z$ [8]. However, the reverse is only true if the distribution is faithful to the graph. That is, if a distribution is faithful to a graph then all independence relationships are represented on the graph and only these relationships are present [9]. Assuming faithfulness allows the topology of a graph to be learned by testing pairs of variables for independence given various conditioning sets.

2.5 Discriminating paths

For the FCI and RFCI algorithms the notion of discriminating paths is needed. A path $\pi = (A, \dots, X, Y, Z)$ is discriminating a discriminating path for Y if:

1. π must include at least 3 edges
2. Y is a non-endpoint on π and is adjacent to Z on π .

3. A is not adjacent to Z and every other node is a collider on π and a parent of Z. [10]

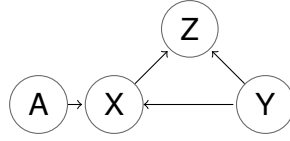


Figure 2.6: $A - X - Y - Z$ is a discriminating path on Y

2.6 Possible D-Separation Set

The FCI algorithm must also find the possible-d-separation set for every node on the graph. This set consists of all nodes which could be d-separated based on some conditioning set. For a particular node (X) in a graph this set consists of all nodes (Y) for which there exists a path from X to Y for which every triple (A,B,C) is a triangle or the node B in the triple is a collider on the path [11].

2.7 Tests for Independence

When trying to find causal relations between variables tests for independence and conditional independence must be performed. These tests allow us to find correlation between variables and also find out if the correlation between two variables is actually caused by another.

One such test is the χ^2 test, this test gives the probability that an observed distribution is equal to an expected distribution.

If the variables X and Y are independent then $p(X, Y) = p(X)p(Y)$. In this case $p(X, Y)$ is the observed distribution and $p(X)p(Y)$ is the expected distribution. Both of these distributions can be calculated from data. The χ^2 test can then be used to find the probability that they are equal [12].

To estimate $p(X, Y)$ the counts of all pairs of values taken by the variables can be found. They can then be divided by the count of all data points. These counts can be stored in a table with the values of X as columns, values of Y as rows (or visa versa) and entries as observed counts, for ease of access. This type of table is known as a contingency table.

To estimate the distribution $p(X)$, the count of points in which X takes a value over the total number of points for all values of X . The counts can be

2 Literature Review

X	1	1	0	1	0	1	1
Y	0	1	1	1	0	1	0

		X	
		0	1
Y	0	1	2
	1	1	3

Table 2.1: Example Contingency Table For Two Variables

found with the sum of the columns of the contingency table. The same can be done for $p(Y)$ with the rows of the table.

The χ^2 statistic can then be calculated with:

$$\chi^2 = \sum \frac{(p(X)p(Y) - p(X,Y))^2}{p(X)p(Y)}$$

The χ^2 value can then be passed to a cumulative density function and a p-value returned. The function used is dependent upon the degrees of freedom. The degrees of freedom are equal to the $(no.rows - 1) * (no.columns - 1)$. The p-value represents how likely it is that the two variables are independent.

The test is similar for conditional independence. The χ^2 statistic now depends on conditional distributions. If $X \perp Y | Z$ $p(X, Y | Z) = p(X | Z)p(Y | Z)$. Therefore:

$$\chi^2 = \sum \frac{(p(X|Z)p(Y|Z) - p(X,Y|Z))^2}{p(X|Z)p(Y|Z)}$$

These values can all be found by calculating a contingency table for X, Y and Z .

This statistic can then be used in the same way as before to calculate a p-value. However, the degrees of freedom are multiplied by the possible values of $Z - 1$.

2.7.1 Algorithms For Learning Causal Networks

2.7.2 Constraint Based Methods

There have been a number of algorithms developed to learn graphical causal representations from data. This project aims to implement up to three constraint based algorithms: The PC algorithm [13], the Fast Causal Inference (FCI) algorithm [10], and the Really Fast Causal Inference (RFCI) Algorithm [10].

Rather than learning an individual DAG the PC algorithm learns a set of DAGs which could represent the causal relationships found in the data. In cases where the direction of an edge is unclear no direction is specified. The output of the algorithm is a partially directed acyclic graph which represents the set DAGs with all possible orientations of undirected edges [13].

The FCI and RFCI algorithms learn partial ancestral graphs (PAGs) which represent markov equivalence class of MAGs that fit the data [10]. Unlike in the PC algorithm if the direction of an edge can not be determined it is assumed that there is some set of latent variables influencing the nodes being tested. Each end of an edge can be one of three things in a PAG, an arrow in every MAG (\rightarrow), a tail in every MAG ($-$) or to represent the possibility of latent variables having a causal relationship with an observed variable (\circ), ends of an edge denoted by $*$ can be any of the specified types. This allows for a richer display of information than a partial DAG [10].

The PAG generated by the FCI algorithm is considered to be less informative than the PAG generated by the RFCI algorithm. This is because the FCI algorithm performs more tests than the RFCI algorithm. These tests give us more confidence in the presence and the orientation of edges however this comes at the cost of speed [10].

2.7.3 Score Based Methods

The above algorithms are all constraint based. There is another class of algorithm which are score based. Rather than finding graphs using the rules that define the networks, score based methods generate random graphs and assign them a score based on how well they fit data.

An example of a scored based method is the graphical lasso algorithm.

2.8 Computing a skeleton

All three algorithms begin by calculating the skeleton of the graph. The skeleton consists of the edges that are present in the partial DAG or PAG however all edges are undirected. The separation set of two variables ($sepSet(x, y)$) contains the smallest set of variables that the pair are conditionally independent on. When calculating the skeleton the separation set of each pair of variables is also recorded as it is needed to orient the edges of the graph [10], [13].

The skeleton is calculated using a series of conditional independence tests. These conditional independence tests are dependent on the type of input data, therefore conditional independence tests should be defined independently of the algorithm. The conditional independence tests must be able to determine if two variables are independent conditioned on some set of other variables from data [13].

To begin the skeleton estimation the fully connected undirected graph of all variables is constructed and then conditional independence tests are performed to determine which edges to remove.

Adjacent pairs of variables are tested for independence conditioned upon a disjoint set of variables adjacent to one of them. The size of these sets starts at 0 but is incremented after all adjacent pairs have been tested. This repeats until there are not enough adjacent variables to fill the conditioning set.

The edges which have not been removed form the skeleton of the graph which is needed for the PC, FCI and RFCI algorithms along with the

recorded separation sets of removed edges [10], [13].

Algorithm 1: Algorithm to Estimate Skeleton of Causal Network [13]

Input: Data, Independence Test

Result: Estimated Skeleton: G , Separating Set of each Variable:
 $sepSet$

begin

$k = 0$

N = list of variables which make up the data

G = fully connected undirected graph with nodes N

while *There is some $S \subseteq adj(X, G) \setminus Y$ and $|S| = k$ for some*
 $X \in N, Y \in adj(X, G)$ **do**

for $X \in N$ **do**

for $Y \in adj(X, G)$ **do**

for $S \subseteq adj(X, G) \setminus Y$ *where $|S| = k$* **do**

if $X \perp\!\!\!\perp Y | S$ **then**

 remove edge $X - Y$ from G

$Sepset(X, Y) = S$

$Sepset(Y, X) = S$

$k+ = 1$

 Return $G, sepSet$

2.9 The PC Algorithm

After generating the skeleton and separation sets of pairs of nodes, the pc algorithm attempts to orient the undirected edges of the graph [13].

It begins by searching for colliders. For each triple of vertices A , B , and C where A and B are adjacent, B and C are adjacent, but A and C are not adjacent orient the edges $A-B-C$ to $A \rightarrow B \leftarrow C$ if B is not in the separation set of the pair (A, C) . [13] This can lead to some colliders being overwritten by other colliders which share an edge making this part of the algorithm order dependent .

Once colliders are specified other edges can be oriented. This is done by repeating two steps until no more edges can be given a direction. The process of edge orientation is described in Algorithm 2.

Algorithm 2: PC Algorithm to Orient Edges [13]

Input: Skeleton of Causal Network, Separation sets of variables

Result: Partial Directed Acyclic Graph

begin

$G = \text{Skeleton}$

$N = \text{nodes of } G$

for $i, j, k \in N$ **do**

if $k \notin \text{adj}(i, G)$ **and** $j \notin \text{sepSet}(i, k)$ **then**
 | orient $i - j - k$ as $i \rightarrow j \leftarrow k$

while *No more edges can be oriented* **do**

for $i, j, k \in N$ **do**

if $k \notin \text{adj}(i, G)$ **then**

 | orient $i \rightarrow j - k$ as $i \rightarrow j \rightarrow k$

if *There is a directed path from i to j in G* **then**

 | orient $i - j$ as $i \rightarrow j$

 Return G

2.10 The FCI algorithm

The FCI algorithm starts in the same way as the PC algorithm using exactly the process of conditional independence tests to estimate a skeleton of the final graph. The skeleton is then converted to a PAG by setting all edges to $\circ - \circ$. It orients colliders in the same way, however instead of having to overwrite edges bidirectional edges can be created by overlapping colliders [11].

Next, the final skeleton of the graph is obtained from the initial skeleton G . To do this The possible d-separators of every variable must be found in G . These sets are not updated when edges are removed so the order of testing does not affect the outcome. All adjacent variables (X, Y) are the tested for conditional independence, conditioned on every subset of each of the possible d separators set. If the test finds conditional independence the edge X, Y is removed and the conditioning set recorded as a separation set [11].

Next the colliders must be reoriented. This is done in the same way as earlier.

Once the final skeleton has been estimated, A series of orientation rules can be performed repeatedly until no more orientations can be found. Theses rules where originally described in. However these rules did not

fully orient the — tags of the graph [14]. 10 rules were found in [14] which allow full orientation of the graph [10]. The complete set of rules is used in the R implementation and will also be used in the python implementation [1].

Algorithm 3: FCI Algorithm to Orient Edges [11]

Input: Skeleton of Causal Network: G , Separation sets of variables:
 $sepSet$

Result: Partial Ancestral Graph

begin

```

     $N$  = nodes of  $G$ 
    Set all edges of  $G$  to  $\circ - \circ$ 
    for  $i, j, k \in N$  do
        if  $k \notin adj(i, G)$  and  $j \notin sepSet(i, k)$  then
            orient  $i * - \circ j \circ - * k$  as  $i \circ \rightarrow j \leftarrow \circ k$ 
    for  $i \in N$  do
        for  $j \in adj(i, G)$  do
            for  $T \subseteq PossibleDSeps(i, G)$  do
                if  $i \perp\!\!\!\perp j | T$  then
                    remove edge  $i * - * j$  from  $G$ 
                     $sepSet(i, j) = T$ 
                     $sepSet(j, i) = T$ 
    Set all edges of  $G$  to  $\circ - \circ$ 
    for  $i, j, k \in N$  do
        if  $k \notin adj(i, G)$  and  $j \notin sepSet(i, k)$  then
            orient  $i * - \circ j \circ - * k$  as  $i \circ \rightarrow j \leftarrow \circ k$ 
    while No more edges can be oriented do
        Apply R1 – R10 from [14]
    Return  $G, sepSet$ 

```

2.11 The RFCI Algorithm

Like the other two algorithms the RFCI begins by estimating the skeleton of the final graph. Next a list of all unshielded triples must be found. The skeleton, separation set and list of triples are fed to Algorithm 4.

Algorithm 4 orients V-structures and updates the separation set. The updated graph is then fed to Algorithm 5 to orient the graph fully.

Algorithm 5 uses many of the rules used in Algorithm 3 to orient edges on the graph. It also utilises Algorithm 4 to update the skeleton of the graph

as edges are oriented [10].

Algorithm 4: RFCI Algorithm to Orient V-Structures [11]

Input: Skeleton of Causal Network: G , Separation sets of variables:

$sepSet$, List of unshielded triples: M

Result: Partial Ancestral Graph

begin

N = nodes of G

L = empty list

while M is non-empty **do**

 choose $\langle i, j, k \rangle$ from M

$T = sepSet(i, k) \setminus j$

if $i \perp\!\!\!\perp j | T$ and $k \perp\!\!\!\perp j | T$ **then**

 add $\langle i, j, k \rangle$ to L

else

for $r \in i, j$ **do**

if $r \perp\!\!\!\perp j | T$ **then**

 Find the minimal separating set $Y \subseteq T$ for r and j

$sepSet(r, j) = Y$

$sepSet(j, r) = Y$

 add all triples $\langle r, *, j \rangle$ to M

 Remove all triples containing r and j and either r or j in the middle from M and L

 Remove $r * - * j$ from G

 Remove $\langle i, j, k \rangle$ from G

for $\langle i, j, k \rangle \in L$ **do**

if $j \notin sepSet(i, k)$ **then**

 Orient $x * - \circ j \circ - * k$ as $x * - \rightarrow j \leftarrow - * k$

 Return $G, sepSet$

Algorithm 5: RFCI Algorithm to Orient Remaining Edges [11]

Input: Partially oriented Causal Network: G , Separation sets of variables: $sepSet$

Result: Partial Ancestral Graph

begin

G = network

N = nodes of G

while no more edges can be oriented **do**

 orient as many edges of G as possible using $R1 - R3$ from [14]

for $l, j, k \in N$ where l, j, k is a triangle **do**

while Edges $j \circ - * k, i \leftarrow * j, l \rightarrow k$ exist **do**

 Find shortest discriminating paths for l, j and k

if One of these paths π is between k and i **then**

for adjacent nodes r and $q \in \pi$ **do**

$l = -1$; **while** $|sepSet(i, k) \setminus \{r, q\}| \geq l$ **do**

$l++$

for $Y \subseteq sepSet(i, k) \setminus \{r, q\}$, where $|Y| = l$

do

if $r \perp\!\!\!\perp q | Y$ **then**

$sepSet(r, q) = Y$

$sepSet(q, r) = Y$

 Create list M of all triples $\langle r, *, q \rangle$ which form a triangle in G

 Remove edge $r * - * q$ from G

 Run Algorithm 4 with $C, sepSet$ and M as input

if All edges in π are present in G **then**

if $j \in sepSet(i, k)$ **then**

 Orient $j \circ - * k$ as $j \rightarrow k$

else

 Orient $l \leftarrow j \circ - * k$ as $l \leftrightarrow j \leftrightarrow k$

 Orient as many edges as possible using $R5 - R10$ from [14]

 Return $G, sepSet$

3 Design

3.1 Methodology

When designing software there are a number of approaches that can be taken. Most of these approaches break development down into a number of different phases and describe the order in which they should be completed.

Most methodologies were designed for teams of engineers, as this project was completed by one individual many components of the methodologies are redundant. However there are some principles and techniques which can be utilised by an individual developer and aid in the organisation of a project.

3.1.1 Waterfall

The waterfall methodology is a linear approach to development. The project is broken down into stages such as design, implementation and testing. Each stage of development is completed sequentially. This means that once a stage is completed there is no need to return to it and the development process flows in one direction like a waterfall [15].

The waterfall approach is good because the whole project is rigorously planned and documented before coding begins. This allows major difficulties in projects to be considered and accounted for. It also allows projects to be easily passed between developers as a new developer can look at the plan of the project. In less linear approaches there may be plans for future parts of software which are not documented and could be lost by the time of implementation.

The downfall of the waterfall approach comes when a project is not a fixed size or there is a change in requirements, there is no stage in which plans can be modified as that would involve going backwards in development. This makes the project hard to modify once certain stages, such as design, have been completed .

3.1.2 Agile

The agile approach consists of a set of principles which focuses on iterative development [16]. Deliverables, or functioning pieces of software, are identified and designed individually, there is also a much greater focus on executable code than documentation compared with waterfall.

Many parts of the agile methodology are not needed in this project such as ways to enable cooperation between developers as this is a solo project.

The most important part of agile is its ability to deal with changing scope throughout the project. Since testing is done at each stage it will allow the project to be easily shortened or extended based on time constraints, as after every sprint a working piece of software has been developed. Therefore so long as each sprint is completed the project will have produced working code. In waterfall this may not be the case.

3.1.3 Spiral

The spiral model for software development is a risk driven process. The spiral method can contain aspects from the other processes. [17]

The actions taken and the amount of work done on areas of the project are determined by what would minimize risk.

Risk can be a number of things. In an industrial setting risk could be how how spending more time on a product and delaying it's release would affect it's sales figures. In a team project, risks could consist of members of the team being unable to work on software unexpectedly because of illness. In essence, risk is anything that would be detrimental to the stakeholders in the project.

Spiral development also requires risks to be assessed at every stage of development. This would increase the work load specifically in agile as risks would need to be identified at every iteration. This would lead to a lot of repeated work that could be better spent on development or testing.

3.1.4 Chosen Approach

The chosen approach for this project was agile. The reduced amount of documentation will speed up development. Agile development also allows the scope of the project to change over time. This allowed the scope of the project to start off small and increase in size as time allowed it.

For example, once the PC algorithm had been implemented, if time

allowed, the project could easily be extended to include the FCI algorithm.

The spiral approach will mostly not be used for this project as it will lead to an unnecessary increase in work for a relatively low risk project. The project also has a fairly logical order of implementation so using risk minimization to determine what to implement will not be necessary. However the choice of agile due to the risk of changing scope could be considered a use of the spiral method.

Within the Agile methodology there are a number of popular frameworks. However, these teams are designed for teams of software engineers with clients. Since this is a solo project without a client many parts of these frameworks will not be applicable. Therefore, the applicable parts of various frameworks were chosen and compiled into a framework that will work for this project.

3.2 Sprints

Sprints are sections of development defined by deliverables, sprints are an aspect of the SCRUM framework [18]. They outline the tasks required to be completed to develop a deliverable and the time frame needed. The time frame for these sprints was substantially longer than it would be in an industrial setting, this is due to project only being worked on part time by a single individual meaning more time would be needed to create reasonable deliverables.

When a sprint is completed, the scope of the project and the requirements can be updated and modified if needed. For this project there were four main deliverables implemented and four sprints to implement them.

The first sprint consisted of implementing the χ^2 conditional independence test. This test was needed before anything else could be properly implemented.

Next, the skeleton learning software was written. The skeleton is needed for all algorithms in this project so is the next logical step from the independence test.

Then the PC algorithm was fully implemented, this algorithm was the simplest of the 3 so made sense to start with. Starting with the easiest algorithm gave me experience using networkx and the other libraries being used that would be very useful in the final sprint.

Finally, the FCI algorithm was implemented. This was more complex than the pc algorithm, however the experience with the library in previous sprints greatly helped with the more complex orientation.

If the project were to continue the next sprint would consist of implementing the RFCI algorithm which would easily follow on from the FCI implementation.

3.3 Requirements

Requirements for the project were elicited at the beginning of a sprint. In each sprint a particular deliverable was chosen and the requirements of that deliverable were found. Requirements were found by considering how the software would be used. This is an agile approach to requirement elicitation.

3.3.1 Epics

Epics describe a very high level interaction between a user and the software. They describe the general functionality of a deliverable without discussing the details of implementation. Epics were the first stage of requirement elicitation as they capture the functionality that the requirements must capture [19].

3.3.2 User Stories

Each epic is then broken down into user stories. The stories describe what a user must be able to do with the software during an epic. They are fairly low level and form a set of required functionality of the software [19].

Users stories can be used as requirements as if a deliverable can fill the role described in a user story it must be functional. So long as the user stories are complete the software implemented based on them will be too.

3.3.3 Tasks

From the stories tasks can be formed. These tasks describe what exactly must be done to fulfil the requirements. Each sprint had a number of associated tasks. The tasks were small and each task could be verified through testing.

Each task has associated stories so that all software can be traced back to an epic. Tracing back to an epic shows that each task will in some way contribute to functional software.

3.4 Architecture

Only the architecture for a particular deliverable was designed in each sprint. This meant that the architecture grew over time. Considerations for general extensibility of the software needed to be made in the design phase to ease design and implementation during later sprints.

A class diagram was produced, this specified the relationships between classes, such as inheritance and which classes would use instantiations of the other classes. The functionality of each class. This diagram evolved throughout the project as various sprints were completed.

The class diagram for this software was fairly straightforward. This is because the algorithms did not require many classes to be interacting with one another. Each algorithm runs in a single thread and user interaction is minimal. This resulted in a small number of classes keeping the project easy to understand for maintenance and future development purposes.

3.4.1 Unit Testing

Test driven development (TDD) was one of the main features of the Extreme Programming development framework used [20]. When using TDD, tests were created before the actual software was written. This ensures that all software that is produced functions correctly. It also ensures that any refactoring performed would not compromise the functionality of software. Changing one piece of code in this scenario affected another piece of code that was using it would be very apparent in TDD, whereas this could be missed in other forms of development.

All software was tested using the python unittest framework. In this framework, list of tests can be compiled into test cases and cases compiled into a test suite. Cases test an individual feature of software, so can be run individually when a feature is being developed or maintained. Test Suites can then test the software as a whole.

3.5 Documentation

All software written was thoroughly documented. Every class, method and function was given an attached docstring detailing its functionality and use.

NumPy style documentation was used [21]. An example is shown in Figure 3.1. This style was chosen because it gives a useful summary of the functionality of a piece of code. It also allows other developers to easily

3 Design

```
""" A function to build the set of conditioning sets to be for variables x and
y on graph g of a certain size when generating a skeleton
Parameters
-----
X : str,
One variable being tested for independence
Y : str
The other variable being tested for independence
graph : float, optional
The minimum p-value returned by the independence test
for the data to be considered independent
size: int
the size of each conditioning set to be returned
Returns
-----
list of lists of strings
a list of conditioning sets to be tested
"""
```

Figure 3.1: Numpy Style docstring

use the code by showing exactly what arguments are taken and exactly what will be returned by a function.

Inline comments were also used to clarify areas of code within methods or functions. These comments can allow a developer to quickly and easily understand the code for the purposes of extension or maintenance. They were not as strict in format as the documentation and were mainly used to clarify the code below them.

3.6 Tools

A number of existing tools were used throughout the project to aid in development.

3.6.1 Version Control

An important part of software development is Version Control (VC). VC allows changes to be made to software without risk of losing previous iterations. At any point software can be reverted to a previous version if needed. VC also helps with agile development as deliverables can be easily identified in a VC system.

Git is a one such VC system. Git was chosen mainly because GitHub provides free repositories which can be used to store code. This makes code accessible from anywhere and greatly reduces the risk of it being lost. Github also allows distribution to other developers around the world making the project easily accessible. Github makes git repositories incredibly easy to set up and maintain.

3.6.2 Python Libraries

Various external python libraries were used to facilitate development. They greatly reduced the amount of code that needed to be written increasing development speed and ease.

Pandas is a library containing DataFrames which allow easy manipulation of data. DataFrames store a number of observations of a particular set of variables making them perfect for this application [22]. There is also a built in method to calculate a contingency table of the data, this is very useful in the χ^2 test of independence.

NetworkX is a library used for the creation and manipulation of graphical models [23], the Graph and Digraph were used and extended to handle various graphs used in the algorithms. The built in functionality of the graphs helped with the implementation of the PDAG and PAG classes. NetworkX also has functions for finding paths in graphs, this was used many times in the various edge orientation rules of the algorithms.

A function from the Scipy statistics library [24] was used to in the χ^2 test to calculate the χ^2 statistic and p-value from lists of observed and expected values.

Finally, Numpy arrays used to store the adjacency matrices of graphs. Numpy allows efficient computation with arrays and is used by many other libraries, including pandas. [25]

3.7 Refactoring

Refactoring is a very important part of the agile framework and allows a developer to apply newly gained knowledge to old code. Refactoring consists of rewriting functioning code to improve it.

This improvement could be one of many things. Obvious things like speed and memory consumption may be optimised. However, refactoring can also aid in making code more readable or using a new library that is better suited to the task.

3 Design

Refactoring may however fail and render previously functional code non-functional. However, this is easily mitigated by unit testing and version control. Unit testing allows a developer to see whether code that has been refactored still functions in the same way as old code, tests will fail if refactoring has been unsuccessful. In the case where refactoring fails a developer can either attempt to identify the problematic area of code and repair/replace it or use the version control system to easily roll back to an existing functional version of the software. These two systems hugely mitigate the risks of refactoring.

4 Implementation

4.1 Independence Test

The first sprint consisted of implementing the χ^2 tests for independence and conditional independence. This test is needed for skeleton estimation, which is used in the PC, FCI, and RFCI algorithms so is the logical starting point of the project. This test will only work on discrete data however other tests could be substituted for other types of data.

Pandas DataFrames were chosen to store data throughout execution as they provide easy indexing, labelling and many useful functions. For example, the crosstab function was used to find contingency tables for the χ^2 test.

A function `prepare_data` was implemented to read data from a file into a data frame for processing. This function is a method of all algorithm classes in the project.

The unconditional independence test was fairly straightforward to implement as it only consisted of computing the contingency tables and then comparing the values in the tables to the sums of the rows and columns in the tables.

The conditional case was more complex, first the conditioning set was combined into a single variable through concatenation of each variable and every data point in the conditioning set. This made making a contingency table across all the observed values significantly easier. Next a contingency table across the three variables was needed. Again the crosstab function was used, however calculating the sums of each time a variable takes a particular value was now more complex, this is because the columns for the conditioning set were split between the columns for the Y variable.

On top of this not all values of Y and Z were observed simultaneously in data so we cant simply iterate through all values of Y and Z and sum up the associated columns as some do not exist. First the columns that exist must be found and then they can be summed over. After this the test is similar to the unconditional however the expected valued will now also account for the value of the conditioning set.

In some cases $p(X, Y|Z) = 0$, this would cause a divide by zero error, to

mitigate this $p(X, Y|Z)$ was replaced with $\max(p(X, Y|Z), 0)$.

The `scipy chi2` function was then used to calculate the χ^2 statistic and p-value which is returned by the `implementations` function.

4.2 Skeleton

4.2.1 Variables to be Tested

With this function the skeleton can be found. A fully connected networkx Graph is generated and then edges between nodes that show conditional independence are removed.

First unconditional independence tests are performed between each pair of variables and edges removed between those that are found to be independent. The conditional tests based on variables adjacent to the first variable with size of the set increasing by one after all tests have been performed. The algorithm tests $X \perp\!\!\!\perp Y|Z$ and $Y \perp\!\!\!\perp X|Z$, this may seem inefficient as it appears that tests are being repeated. However, the members of the set Z are dependent on the first variable in the statement. This means that the tests must all be performed.

To find the conditioning sets, the `combinations` function from the python library `itertools` was used to find all subsets of $adj(x, G)$ of a certain size.

4.2.2 Order Dependence

The members of set the conditioning set Z for the test $X \perp\!\!\!\perp Y|Z$ depend on the tests performed before it. This is due to Z consisting of nodes adjacent X , however the set of nodes adjacent to X changes as edges are removed.

In the R implementation an order independent version has been developed. When the size of conditioning sets is incremented, the conditioning sets for all pairs of variables are calculated and stored. However the original order dependent version can still be used in the R implementation.

4.2.3 Separation Sets

Edge orientation requires the separation set of each pair of variables. This was stored as a dictionary of all pairs of edges and simply updated with a conditioning set if that set was found to condition independence between the pair.

Since all of the algorithms learn a skeleton, the function is a method of the GraphLearner class which is the parent class of the other algorithm classes. The function returns a networkx Graph containing the estimated edges of the skeleton and a dictionary containing the separation set of each pair of variables.

4.3 PC Edge Orientation

For the PC algorithm edge orientation is relatively simple. However, the PC algorithm outputs a PDAG which contains a mixture of directed and undirected edges, networkx does not contain a data structure to support this kind of graph so one had to be implemented. This turned out to be fairly trivial, using a directed graph behind the scenes in which undirected edges were represented as bidirectional edge. Since a DAG there cant contain any bidirectional edges so all bidirectional edges can simply be considered to be undirected.

When orienting edges of a graph, one graph was used to store directed and one used to store undirected. When an edge was directed it was simply transferred from one to the other and once all edges that can be oriented are oriented, the two graphs are combined into a PDAG.

4.3.1 V-Structures

The first part of edge orientation consists of identifying colliders. This is fairly simple as you simply need to iterate over all triples, test three edges and check the separation set.

However, there may be situations in which the order of the v-structures you test affects the final orientation of the graph. There may be two connected v-structures where orienting one will also orient the other but not in the direction described by the constraints.

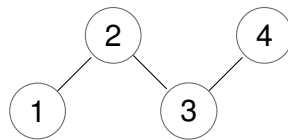


Figure 4.1: A skeleton of a Causal Network

Figure 4.1 shows two possible orientations of the skeleton in Figure 4.2, depending on which v structure is oriented first, in the first case when we arrive at the 2,3,4 triple one of two thing can be done. The edge (3,4)

4 Implementation

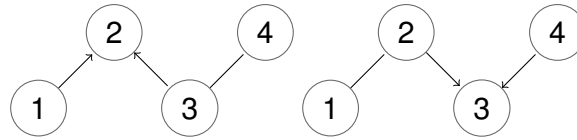


Figure 4.2: Two potential colliders in a skeleton

can be left undirected or the new structure can override the old one, as shown in Figure 4.3. In the pre-existing R implementation the second

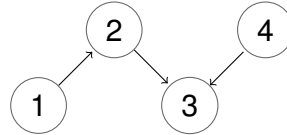


Figure 4.3: A potential orientation of the skeleton in Figure 4.1 containing colliders from Figure 4.2

approach of overwriting edges was taken so the same was done in the python implementation for consistency [11].

Finally the other orientation rules described in the algorithm were followed, these were easy to implement using the networkx graphs. The `all_simple_paths` functions was very useful in the second orientation rule.

4.4 FCI Algorithm Edge Orientation

The FCI algorithm was able to use the same skeleton estimation method, the collider orienting code was also very similar. However after this code unique to the FCI algorithm must be developed.

First the final skeleton of the graph must be found. This is done by performing more independence tests, conditioned on the possible d-separating sets of the variables. Finding these sets was made simple by a networkx function to find paths between nodes on a graph.

For the rest of the algorithm the network is a PAG. The data structure used for the PAG was a subclass of the networkx undirected graph. In this graph the direction of an edge was determined by tags stored in a dictionary associated with that edge. This allowed all edge types of a PAG to be stored.

4.5 Testing

To ensure that the algorithm was behaving as intended all parts must be thoroughly tested. The python unittest module was used to create a test suite covering as many aspects of the algorithm as possible.

4.5.1 Independence Test

For the χ^2 independence test, a sanity check consisting of testing whether a variable was independent of itself was used. Obviously the test will result in a p-value of 0 so was very easy to test. This test was repeated in the case of a conditioning set of size 1 and size 2 to ensure that it's did not affect functionality.

Next a pre-existing test from the R bnlearn library was used to find results on real data. This data was tested with the python implementation and the results compared. Data from the Alarm_10000 the Asia_10000 and the Asia_1000 datasets were used from [26]. These datasets have been generated from casual networks and are discrete, this makes them good to test software which estimates these graphs and takes discrete data. Many tests with different variables and conditioning sets were performed.

4.5.2 Skeleton Estimation

To test the skeleton estimation again the R implementation was used. However, the R version does not use a chi squared test, a wrapper needed to be written to make it compatible with the bnlearn function. With this, skeletons could be generated from data in the same way in python and R and the results compared. The separation sets were also recorded and compared.

4.5.3 PC Algorithm

Testing the edge orientation for the PC algorithm first started with small sanity checks. For example the collider orientation was tested on networks with 3 nodes and two edges and simple separation sets.

The path checking needed for the PC algorithm was also tested on small networks where paths were linear and some with many branches. It was also tested on a few negative examples.

Finally a skeleton and separation set generated from the alarms data was

oriented in both R and python to test the whole edge orientation process as a whole.

4.5.4 FCI Algorithm

For the FCI algorithm, the first part of the skeleton estimation is the same as in the PC algorithm. However, the final skeleton estimation needs to be tested so comparison testing was done with the R implementation.

Comparison of the edge orientation was performed, however due to the number of edge orientation rules it is likely that only some of them would be applied in a particular run. This means that to fully test the rules they should be tested individually. So the rules were tested on graphs which the rule should be applied to and graphs which the rule should not be applied to. The graphs were then tested after rule had been applied to see if the edges had been oriented correctly.

4.5.5 Final Structure

The final implementation consisted of four main components. The χ^2 test for independence, the skeleton estimation, PC algorithm, and FCI algorithm.

4.5.6 Independence Test

The χ^2 test was a stand-alone function that can take a data frame along with the names of variables to be tested and a conditioning set. It returns a the χ^2 statistic and a p-value from the test. The p-value represents the probability that the variables are not dependent conditioned on the conditioning set, if this statistic is high enough the variables can be considered independent.

4.5.7 Graph Learners

Each algorithm is implemented as a class, the base class of all algorithms is the graph learner class. These classes take data and an independence test and learn a graph.

The base class implements the skeleton learning method that all the algorithms use. It can also find and orient colliders on that skeleton. Finally, it has a static method to prepare data, this reads data from a file into a dataframe which can be used by the other methods.

4 Implementation

The PC algorithm and the FCI algorithms were implemented as sub-classes of the Graph learner. The PC algorithm only needed to implement the edge orientation step to complete the algorithm.

The FCI algorithm had to both implement edge orientation and extend the skeleton estimation of the graph learner.

4.5.8 Graphs

Two new graph classes were implemented, the PDAG and the PAG. The PDAG is a fairly simple extension of networkx's built in Directed graph class in which undirected edges are merely stored as bidirectional edges, in PDAGs there are no bidirectional edges so there will be no confusions between bidirectional and undirected edges.

This implementation is simple but fairly powerful. Most built in networkx functions will work in a reasonable manner. For example, the `has_edge` function will be direction dependent on directed edges but not on undirected edges. Path finding algorithms will also work assuming undirected edges can be traversed in either direction.

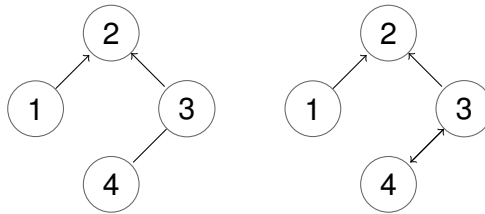


Figure 4.4: A PDAG and how it would be represented in the implementation

The PAG was more complex to implement due to the increased complexity of the graph. The PAG class is an extension of the undirected graph from networkx. All information about the tags, and therefore the directionality of the edges is stored in an attribute associated with particular edge. The attribute is a dictionary with keys being the nodes at each end of the edge and the values being the tag at that end of the edge.

This implementation meant that many new methods needed to be added and many existing methods modified. For example, when adding a new edge the tags of the edge needed to be assigned. methods for testing if edges were directed from u to v (had an arrow head into v) or fully directed from u to v (tail at u , arrowhead at v) were created and used during the FCI algorithm.

The built in methods for finding paths did however still work with this implementation which was very useful for edge orientation.

4 Implementation

Currently viewing the graph as a whole is tricky, the built in functions for displaying graphs can not display the edge tags. It may be possible to use the display edge labels function to show the tags as a label on the graph.

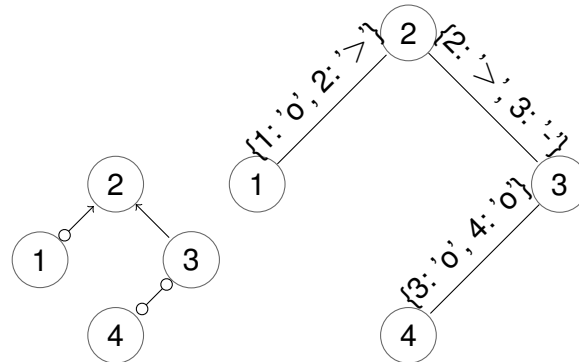


Figure 4.5: A PAG and how it would be represented in the implementation

4.6 Usage

There are currently two main ways to use the python implementation, as a library and as a command line interface.

The usage of the various objects and data structures is described in the doctring of each class, they can simply be imported and used as part of a wider program.

To use the PC or FCI algorithm, simply run the program from a command line with the path to the data file as the first argument. Optional parameters are available to state if the data is labelled (default False) and what the deliniator (default ' ') is. Data should consist of rows of observations and columns of variables, the first row may consist of labels for each column.

4.7 Viewing Graphs

There are a number of options to view a learned graph. Built in functions of the parent classes allow the edges and nodes to be iterated over to see where causal links have been estimated.

4.7.1 Adjacency Matrices

Options to view the graphs as adjacency matrices have also been implemented. In these matrices the rows represent the from node and the columns represent a to node. The order of the rows and columns is the same as the order in which the nodes are iterated over. The matrix is stored as a numpy array so it can be used with efficient numpy methods in the future.

In the PDAG when there is a directed edge from one node to another there is a one in the appropriate position. If there is not a directed edge there is a zero. Directed edges going in both directions between nodes represent an undirected edge.

The PAG uses numbers greater than zero to represent the tags on an edge and zeroes to represent the absence of an edge. the '0' tag is represented as a one, the '-' tag as a two and the '>' tag as a three.

These matrices can also be written to files for future use. Each row of the matrix is written as a row values in a row are separated by spaces. The first row of the file consists of the node name of the respective column. The first value of each row of points is the name of the corresponding node. This representation can easily be used with a number of libraries including pandas, it can also be easily loaded into an R environment as a dataset.

5 Analysis

Analysis of this project will mainly consist of comparison to the implementation of the algorithms in R, where possible. It will compare the functionality of the algorithms on data and also the run speed of the code.

5.1 Initial Skeleton Generation

The skeleton generation code has almost identical functionality as the R implementation. Even the order in which edges are tested is the same. However the R implementation has a non-order dependent mode which is not currently implemented in the python version however could be added as an option in the future.

The Alarm_1000, Asia_1000, and Diabetes_1000 datasets found in all produce identical skeletons.

The χ^2 test for independence was not implemented in the pcalg library in R. Other independence test for discrete data, such as the G^2 test, were implemented. However, the bnlearn library has an implementation of the χ^2 test to use for comparison.

Both tests return identical results on all tested data as seen in

5.2 PC Algorithm Edge Orientation

There are some differences in the edge orientation of the different implementations. For example, in the orient of Asia_1000 network, the edge either-bronc is oriented in opposite directions by the two implementations. When tested on other datasets very similar results were found, with identical skeletons but occasional slight differences in edge orientation due to the order in which colliders are oriented. This is due to the order in which the colliders were oriented in the graph. In the R implementation the triple lung, bronc, either is oriented followed by the triple tub, either bronc. This causes the edge either-bronc to be overwritten and reoriented. In the python implementation this is done in the opposite order resulting in a different network.

5 Analysis

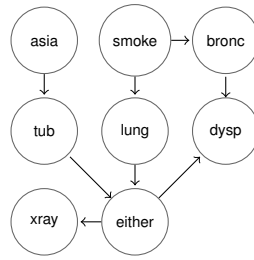


Figure 5.1: The actual network that generated data in the asia_1000 dataset

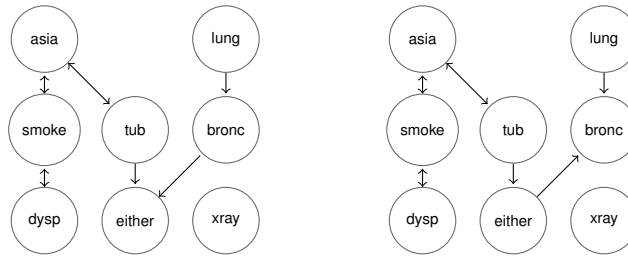


Figure 5.2: The networks estimated by the R (left) and python (right) implementation from the Asia_1000 dataset

Neither order is more correct than the other however this orientation is used in the FCI algorithm so differences in the order can have a knock on effect in the PAGs estimated by this algorithm.

It is worth noting that both estimates are quite different from the actual network which generated the data.

5.3 FCI Algorithm Final Skeleton

The FCI algorithm begins with the skeleton estimation and collider orientation to that of the PC algorithm. However because an edge in a PDAG can be bidirectional the order of of collider orientation will have no effect on the final structure of the skeleton.

To estimate the final skeleton the possible-d-sep sets must be found for every node in the graph. The implementations however find slightly different possible-d-sep sets for the same graph. The R implementation appears to be using a slightly different method to find these sets. This can be illustrated with the Asia_1000 dataset.

The partially learned graph when the possible-d-separation sets are found is the same in both implementations. However the sets found for the 'tub'

node and the 'smoke' node are not equal. The R implementation does not add the 'smoke' node to the possible-d-separation set of 'tub' despite the only node between them being a collider.

This does not affect the estimation of this particular graph, however it will affect other graphs and since it is done fairly early on in the algorithm it can have large knock-on affects on the end result.

5.4 FCI Algorithm Edge Orientation

It would be incredibly difficult to compare how the two implementations apply the rules to orient the final graph. This is due to the above differences affecting the input to this stage. Since all tests of the rules pass and the test are thorough, it is fairly safe to assume that the orientation is valid on a graph without direct comparison to the R version.

5.5 Run Speed

The cProfiling library was used to collect statistics about the implementation. The library collects the number of times each function is called by the code along with the execution time of the code. This profiling allows the bottle necks of the code to be found and made high priorities or optimisation.

The independence test was the main bottleneck of the algorithms, and was significantly slower than the R implementation. This is because the R implementation actually runs C code behind the scenes, C code tends to run faster than python and R code [27], [28].

Calculating the contingency tables was the most expensive part of the independence test (86% of total runtime of the PC algorithm). Currently it uses a built in function from the pandas library. Replacing this function with something faster could make a large difference to the run speed.

However, the independence test in the algorithm is a parameter and thus the current slow test could be wholly replaced by a faster version. This new test could call C code like the R implementation.

6 Conclusion

In conclusion, the project was successful. Two algorithms were implemented which are able to learn causal networks in python. These implementations provide an easy to use framework and a framework that will also be easy to extend in the future.

6.1 Further Work

The most obvious extension to this project would be to implement the RFCI algorithm described in the literature review. Due to time constraints it was not implemented for this project. However the PAG class implements the majority of functionality needed to implement the RFCI algorithm. The process of implementation would be very similar to the process of implementing the FCI algorithm. It was be reasonable to simply extend the FCIAlg class to take advantage of the existing functionality.

Another improvement would be to increase the speed of the conditional independence tests. One approach could be to implement the independence test in C and interface with it using Cython [29]. C programs can be much faster than those written in python so there would almost definitely be an improvement [27].

Other independence tests could also be written. These tests could be for the same kind of data as the χ^2 or it could be a test for other data for example continuous data. The raw data would not even be necessary input to the algorithm, instead some sufficient statistics placed in a pandas data frame could be used if the independence test was able to estimate independence based on these statistics.

A Some apendix

Bibliography

- [1] M. Kalisch, A. Hauser and M. Maechler, *Methods for graphical models and causal inference [r package pcalg version 2.6-0]*. [Online]. Available: <https://cran.r-project.org/web/packages/pcalg/index.html>.
- [2] G. Piatetsky, *Kdnuggets*. [Online]. Available: <https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html>.
- [3] S. O'Grady, *The redmonk programming language rankings: January 2019*, Mar. 2019. [Online]. Available: <https://redmonk.com/sogradyl/2019/03/20/language-rankings-1-19/>.
- [4] T. Verma and J. Pearl, 'Causal networks: Semantics and expressiveness', in *Machine Intelligence and Pattern Recognition*, vol. 9, Elsevier, 1990, pp. 69–76.
- [5] J. Zhang, 'Causal reasoning with ancestral graphs', *Journal of Machine Learning Research*, vol. 9, no. Jul, pp. 1437–1474, 2008.
- [6] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. DOI: 10.1017/CBO9780511804779.
- [7] J. Pearl, 'Causality: Models, reasoning, and inference', *Econometric Theory*, vol. 19, no. 675-685, p. 46, 2003.
- [8] —, 'Causal inference in statistics: An overview', *Statist. Surv.*, vol. 3, pp. 96–146, 2009. DOI: 10.1214/09-SS057. [Online]. Available: <https://doi.org/10.1214/09-SS057>.
- [9] R. Scheines, 'An introduction to causal inference',
- [10] D. Colombo, M. H. Maathuis, M. Kalisch and T. S. Richardson, 'Learning high-dimensional directed acyclic graphs with latent and selection variables', *The Annals of Statistics*, pp. 294–321, 2012.
- [11] —, '(). supplement to "learning high-dimensional directed acyclic graphs with latent and selection variables', *The Annals of Statistics*, 2012.

Bibliography

- [12] K. P. F.R.S., 'X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900. DOI: 10.1080/14786440009463897. eprint: <https://doi.org/10.1080/14786440009463897>. [Online]. Available: <https://doi.org/10.1080/14786440009463897>.
- [13] P. Spirtes and C. Glymour, 'An algorithm for fast recovery of sparse causal graphs', *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.
- [14] J. Zhang, 'On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias', *Artificial Intelligence*, vol. 172, no. 16, pp. 1873–1896, 2008, ISSN: 0004-3702. DOI: <https://doi.org/10.1016/j.artint.2008.08.001>. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0004370208001008>.
- [15] W. W. Royce, 'Managing the development of large software systems: Concepts and techniques', in *Proceedings of the 9th international conference on Software Engineering*, IEEE Computer Society Press, 1987, pp. 328–338.
- [16] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, S. Mellor, K. Schwaber, J. Sutherland and D. Thomas, *Manifesto for agile software development*, 2001. [Online]. Available: <http://www.agilemanifesto.org/>.
- [17] B. W. Boehm, 'A spiral model of software development and enhancement', *Computer*, no. 5, pp. 61–72, 1988.
- [18] K. Schwaber, 'Scrum development process', in *Business object design and implementation*, Springer, 1997, pp. 117–134.
- [19] M. Cohn, *User stories applied: For agile software development*. Addison-Wesley Professional, 2004.
- [20] D. Janzen and H. Saiedian, 'Test-driven development concepts, taxonomy, and future direction', *Computer*, vol. 38, no. 9, pp. 43–50, 2005.
- [21] *Example numpy style python docstrings*. [Online]. Available: https://sphinxcontrib-napoleon.readthedocs.io/en/latest/example_numpy.html.
- [22] *Python data analysis library*. [Online]. Available: <https://pandas.pydata.org/>.
- [23] *Networkx*. [Online]. Available: <https://networkx.github.io/>.

Bibliography

- [24] *Statistical functions (scipy.stats)*. [Online]. Available: <https://docs.scipy.org/doc/scipy/reference/stats.html>.
- [25] *Numpy*. [Online]. Available: <http://www.numpy.org/>.
- [26] [Online]. Available: <https://www.cs.york.ac.uk/aig/sw/gobnilp/data/>.
- [27] I. M. Wilbers, H. P. Langtangen and Å. Ødegård, 'Using cython to speed up numerical python programs', *Proceedings of MekIT*, vol. 9, pp. 495–512, 2009.
- [28] D. j. Wilkinson, *Gibbs sampler in various languages (revisited)*, Nov. 2014. [Online]. Available: <https://darrenjw.wordpress.com/2011/07/16/gibbs-sampler-in-various-languages-revisited/>.
- [29] *C-extensions for python*. [Online]. Available: <https://cython.org/>.