Causal Discovery for Risk Factors and Disease

Yujia Gu

Institute of Statistics and Big Data
Renmin University of China

April 24,2022

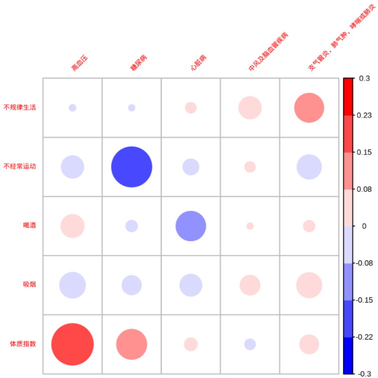
Content

Data Descirption

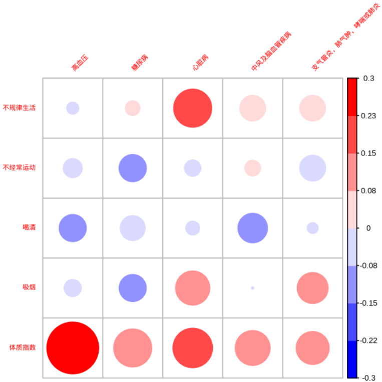
- **Background**: A cross-sectional dataset aiming to study the ageing process was collected by several hospitals using questionnaires and fitness tests, including over 15000 people of different ages.
- **Questions**:
  1. Is there any causal relationship between risk factors and noncommunicate diseases?
  2. If such causation exists, how it changes through the ageing process?
- **Risk Factors**: Drinking, Smoking, Lack of exercise, Irregular lifestyle, Body Mass Index(BMI).
- **Disease**: Hypertension, Heart disease, Diabetes, Apoplexy, Respiratory system diseases

Brief Review on EDA

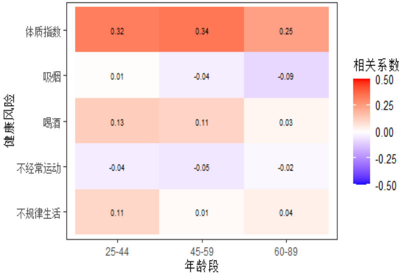
Overall Correlation



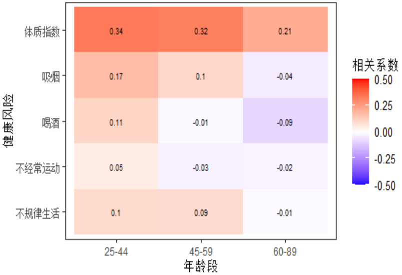
(a) Male



(b) Female



(a) Male



(b) Female

Idea: Using directed acyclic graphs(DAG) to model the joint distribution of all variables and discover the causal relationship between them. We will write the joint distribution as

$$\Pr(X_1, \ldots, X_m) = \prod_{i=1}^{m} \Pr(X_i | pa(X_i))$$

Example:

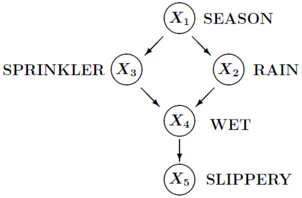


Figure 1: A Bayesian network representing causal influences among five variables.

Assumptions

Local Markov Condition: Every X_i in DAG is independent of its non-descendants conditional on its parents(d-separation property).
Faithfulness Assumption: DAG demonstrates all conditional independence relations for the population probability distribution.

Troubles caused by Assumptions

- Different DAGs may share same Markov conditions. Such collections for those DAGs are called Markov Equivalent Class.
- By Faithfulness, confounders are not allowed in the model.

Structure Learning Methods

- Constrained-Based Methods
- Score-Based Methods
- Hybrid Methods
- Functional Causal Models

Constrained-Based

Ideas:

1. Construct undirected graph by independence tests and conditional independence tests.
2. Directions can be constructed by v-structure.

Algorithms:

- PC algorithm (Spirtes, 2001; Implemented by Colombo,2014)
- iAMB algorithm(Yaramakala,2005)
- FCI algorithm(Spirtes, 2001)

Advantages:

- Convenient interpretation for causal relationships.

Disadvantages:

- Multiple Testing.
- May exist undirected edges.

Score-Based

Ideas:
Each candidate DAG is assigned a network score reflecting its goodness of fit, which the algorithm then attempts to maximise.
Scores:

- AIC,BIC
- NML, fNML,qNML(Shtarkov, 1987; Silander et al., 2008; Silander et al., 2018)

Advantages:

- An optimized DAG which fits the distribution of the data will be produced.

Disadvantages:

- Hard to give explanation on causal relationship.

Hybrid Method

Ideas:

1. Use constrained-based methods to find the skeleton(undirected graph) of the Bayesian network.
2. Use score-based methods to direct the edges.

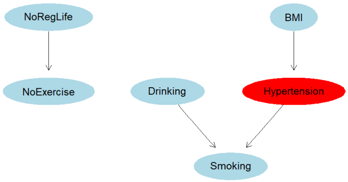
Algorithm:

- MMHC (Tsamardinos et al.,2006)
- Rsmax2 (Scutari et al. 2014)
- H2PC (M. Gasse et al.2014)
- FRITL (Chen et al. 2021)

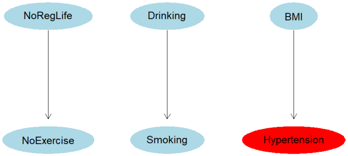
Our Model

- Hybrid method is used for formal modeling.
- Constrained-based and score-based methods are used for comparison.
- R package bnlearn is used for modeling.

Risk Factors v.s. Hypertension



(a) Male



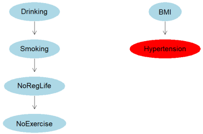
(b) Female

Figure: Risk Factors and Hypertension across all age groups

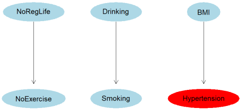
Risk Factors v.s. Hypertension

Different Age Groups

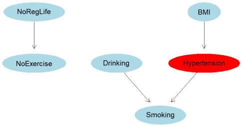
Figure: Male



(a) 25-44

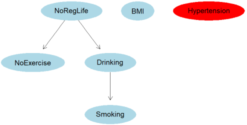


(b) 45-59

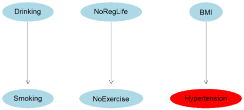


(c) 60-89

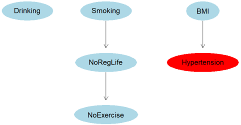
Figure: Female



(a) 25-44



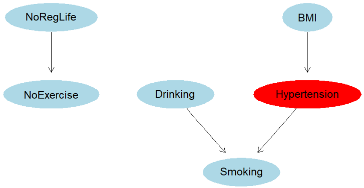
(b) 45-59



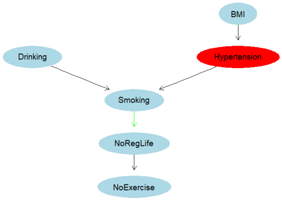
(c) 60-89

Comparison Between Methods

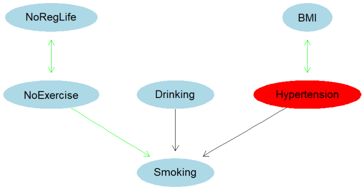
Visual Comparison



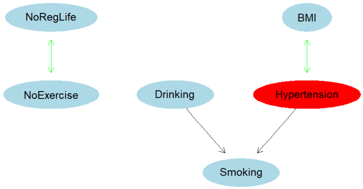
(a) Hybrid



(b) Score-based,qnml



(c) Constrained-based, PC, $\alpha = 0.05$



(d) Constrained-based, PC,$\alpha = 0.001$

5-fold cross-validation was used to compute the log-likelihood loss.

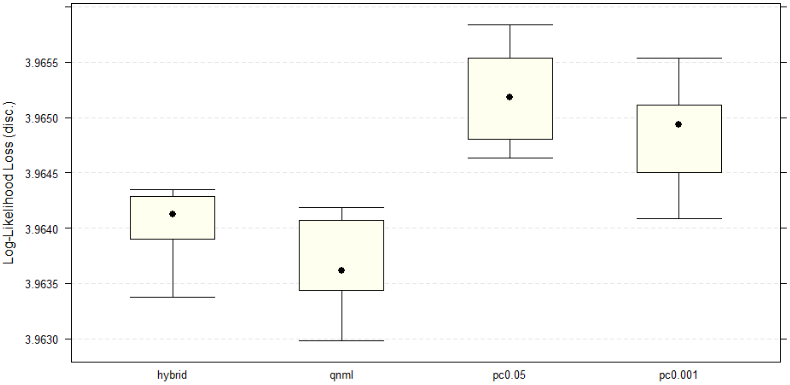
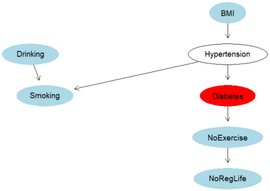


Figure: Log-likelihood Loss Comparison

Confounder

(a) Diabetes without confounder



(b) Diabetes with Hypertension as confounder

Discussion

- Different patterns were shown between genders and age groups.
- BMI may be a causal for hypertension, which meets our common sense.
- For male, hypertension may have influence on smoking. However, people may care more about the reverse direction. At least, our result showed that they have relationship between each other.

Future Work

- Confounders should be considered, using FCI to fit such models.
- We can not figure out the influence is positive or negative by causal discovery. Further works may be focused on this subject.

The End

Questions? Comments?