RestrictedSCMs

June 23, 2019

1 Restricted Structural Causal Models

by Jonas Peters, Niklas Pfister and Rune Christiansnen, 20.06.2019

This notebook is intended to give you some insight on identifiability of additive noise models and how they could be estimated in practice. The references are sparse and are somewhat biased towards [1]. We do not intend to claim that this is the original reference.

```
In [13]: library(mgcv)
         library(dHSIC)
         source("utils.R")
```

1.1 Two Variables

Assume we are given a sample from a bivaraite distribution over X and Y and want to find out, which variable is the cause and which the effect, i.e., whether the distribution has been induced by an SCM with graph $X \to Y$ or $Y \to X$.

Without further assumptions, this will not be possible: Either graph can induce any distribution (e.g., see Proposition 4.1 in [1]). This is different, however, if we restrict the model class. E.g., if we consider only linear assignments, we have the following statement:

Given a distribution over X and Y that is induced by an SCM with linear assignments and graph $X \to Y$, then it is also induced by an SCM with linear assignments and graph $X \leftarrow Y$ only if the noise variables are Gaussian (e.g., see [2] or Thm 4.2. in [1]).

1.1.1 Exercise 1:

Generate a (large) sample from a distribution from a linear SCM with graph $X \to Y$ and non-Gaussian noise (e.g., uniform). Do the same for a linear SCM with the reversed graph $X \leftarrow Y$. Plot both samples. What is the (systematic) difference between the two pictures?

1.1.2 Solution 1:

1.1.3 End of Solution 1

A very similar statement holds for nonlinear functions, too: If a distribution is induced by $Y = f(X) + N_Y$ with N_Y independent of X, then only for very few (and somewhat non-generic) combinations of functions f and distributions of X and N_Y, will we find a model in the backward direction, too (e.g., see Thm 4.5 in [1]).

But how can we find the graph if we are only given a sample from the joint (observational) distribution? In other words, how do we decide which model the data come from:

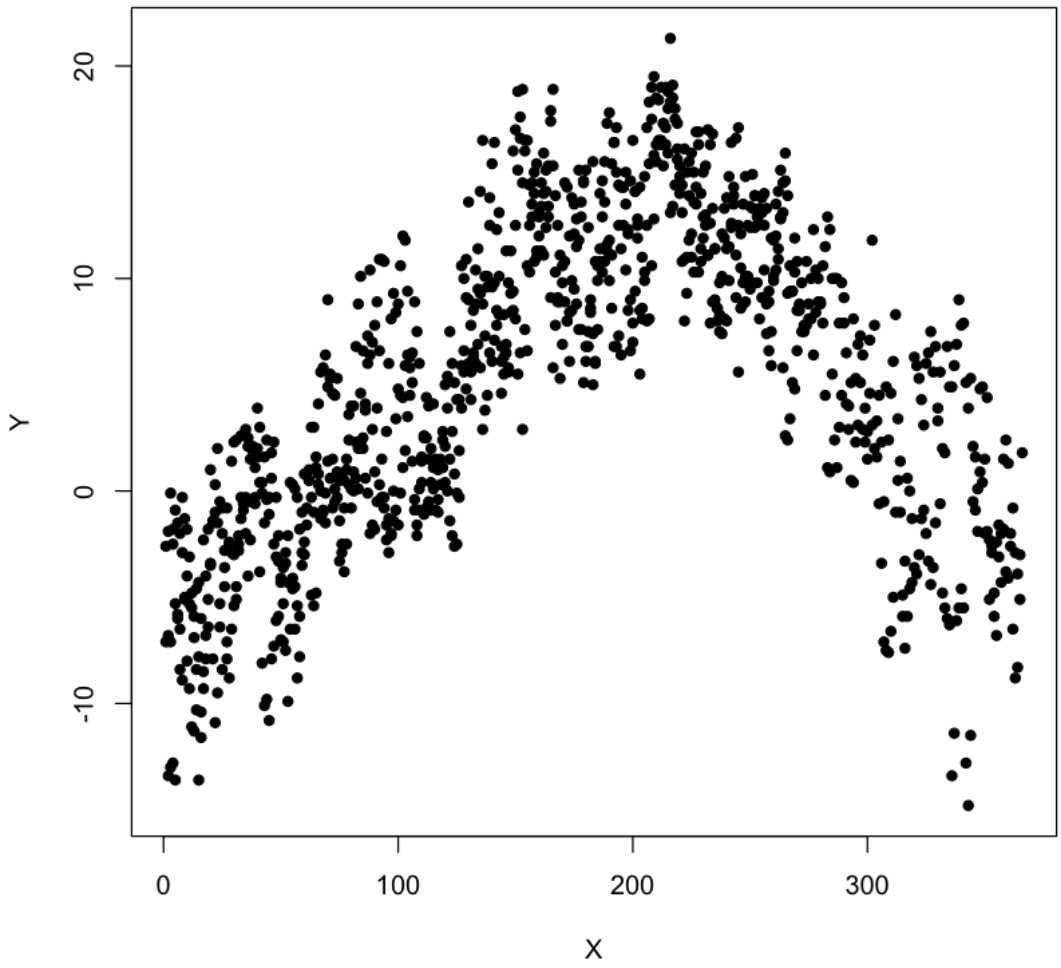
$$Y = f(X) + N_Y, \quad N_Y \text{ indep. } X \tag{1}$$

or

$$X = g(Y) + N_X, \quad N_X \text{ indep. } Y? \tag{2}$$

Let us consider a real world data set.

```
In [14]: XX <- read.csv('./RestrictedSCMsData1.txt', sep = "\t")
         XX <- XX[1:1000,]
         Y <- XX[,2]
         X <- XX[,1]
         plot(X, Y, pch = 19, cex = .8)
```



The next two exercises contain two approaches that aim to estimate the underlying causal DAG, i.e., whether $X \to Y$ or $X \leftarrow Y$.

1.1.4 Exercise 2:

First, we perform an approach based on independence tests. If (1) is correct, then $Y - f(X)$ is independent of X. If (1) is incorrect, then $Y - f(X)$ is dependent on X for all choices of f. We can thus verify (1) by estimating the regression of Y on X and checking if the residuals are independent of X. More formally, we follow the following procedure.

1. Regress Y on X and obtain estimate $\hat{f}$.
2. Check whether the residuals $Y - \hat{f}(X)$ are independent of X.
3. Repeat the same in the opposite direction.
4. If the independence is accepted for one direction and rejected for the other, infer the former one as the causal direction.

To regress B on A, you may use `gam(B  s(A))` from package `mgcv`; the residuals can be accessed by `gam(B  s(A))$residuals`. For the independence test you can use `dhsic.test` from package `dHSIC` (setting `method = "gamma"` will increase the speed).

1.1.5 Solution 2:

1.1.6 End of Solution 2

1.1.7 Exercise 3:

If we know the distribution of the noise variables, we can also choose a maximum likelihood (ML), also called score-based approach. (This means --- roughly--- that we are choosing the model class that is closer to the observed empirical distribution in KL distance.) Given a data set, each possible DAG over variables (X, Y) induces a different likelihood function, different maximum likelihood estimates, and therefore a different likelihood score. We can then choose the graph that obtains the highest score.

Assume that all noise variables are Gaussian with zero mean and a certain variance σ^2. In this case, it can be shown that the likelihood score for DAG $\mathcal{G}$ is proportional to

$$-\log(\text{var}(R_X^{\mathcal{G}})) - \log(\text{var}(R_Y^{\mathcal{G}})),$$

where $R_X^{\mathcal{G}}$ and $R_Y^{\mathcal{G}}$ are the residuals obtained from regressing X and Y on their parents in $\mathcal{G}$, respectively.

For any given graph $\mathcal{G}$, we obtain thus obtain a score propotional to the likelihood score by the following three steps.

1. Regress each node on its parents.
2. Compute the variance of all residuals (here: $R_X^{\mathcal{G}}$ and $R_Y^{\mathcal{G}}$).
3. Calculate the above score.

Compute this score for all graphs (here: $X \to Y$, $X \leftarrow Y$ and the empty graph $X \quad Y$) and compare.