Compute Interventions

by Jonas Peters, Niklas Pfister, 01.12.2018

This notebook is intended to give you some insight on how to compute intervention distributions. We will have a look at two examples.

```
In [ ]: library(igraph) #comment out this line if you cannot install igraph
        library(CondIndTests)
        library(ppcor)
        library(dHSIC)
        source("./utils.R")
```

Given the observational distribution and the graph, we can now compute interventional distributions. As an example, let us focus on the intervention $$ do(X:= 3). $$ This intervention distribution can be computed using parent adjustment, which is also called the G-formula. It says (in case of discrete variables, the integral is replaced by a sum) \begin{equation} \tag{1} p^{do(X:=x)} (y) = \int_{\mathbf{z}} p(y|x, \mathbf{z}) p(\mathbf{z}) \, d\mathbf{z}, \end{equation} where $\mathbf{Z}$ is a ``valid adjustment set''. Usually, there are several valid adjustment sets. One set $\mathbf{Z}$ that always works, is the set of parents of X, i.e., $\mathbf{Z} = \mathbf{PA}_X$.

We will apply formula (1) to a (famous) kidney stone data set. Let us assume that the causal structure looks as follows. \begin{align} &\\ &\begin{array}{cc} &S & \\ \swarrow & & \searrow\\ & & \\ T & \longrightarrow & R\\ \end{array}\\ & \end{align} where T is treatment, S is size of the stone, and R is recovery. The data look as follows
$$ \begin{array}{r|c|c} & \text{Treatment A} & \text{Treatment B}\\\hline \text{Small Stones } { (\frac{357}{700} = 0.51)}& \frac{81}{87} = {0.93} & \frac{234}{270} = 0.87\\\hline \text{Large Stones } { (\frac{343}{700} = 0.49)} & \frac{192}{263} = {0.73} & \frac{55}{80} = 0.69\\\hline & \frac{273}{350} = 0.78 & \frac{289}{350} = {0.83} \end{array} $$

Exercise 1:

Compute $$ P^{do(T:=A)} (R = 1) $$ and $$ P^{do(T:=B)} (R = 1). $$ Which treatment is better?

Solution 1:

End of Solution 1

For the remainder of this notebook we will have a look at an artificial example. We first load some data.

```
In [ ]: load(file = "./ComputeInterventionsData1.Rdata")
```

These data have been generated from an SCM with the following graph structure.

In [ ]:

```
Adj <- rbind(c(0,0,0,1,0,0,0,0,0), c(0,0,1,1,0,0,0,0,0), c(0,0,0,0,0,0,0,1,0),
c(0,0,0,0,1,1,0,0,0),
           c(0,0,0,0,0,0,0,0,0), c(0,0,0,0,0,0,1,1,0), c(0,0,0,0,0,0,0,0,0), c(
0,0,0,0,0,0,0,0,1),
           c(0,0,0,0,0,0,0,0,0))
set.seed(1)
plotGraphfromAdj(Adj, labels = c("C", "A", "K", "X", "F", "D", "G", "Y", "H"))
#comment out the above line if you cannot install igraph
```

Let us say, we are interested in the causal effect from X to Y, i.e., in the intervention distribution $p^{do(X:=x)} (y)$.

Computing equation (1) is often difficult, even if we are given full knowledge of the observational distribution, that is, if we are given the density p (especially in the case of continuous variables and high-dimensional $\mathbf{Z}$). In this example, we are not even given p but only a sample from p and estimating conditional distributions is a hard statistical problem. Fortunately, the joint distribution p is Gaussian and things become easier. Equation (1) then implies \begin{equation} \tag{2} \mathbf{E}^{do(X:=x)} [Y] = \alpha x, \end{equation} where α is sometimes called the causal effect from X to Y and is determined by $$ \mathbf{E} [Y\,| \,X=x, \mathbf{Z} = \mathbf{z}] = \alpha x + \beta^t \mathbf{z} $$ (If you are interested, try to prove that statement.) In practice, we can obtain α as the regression coefficient, when linearly regressing Y on X and $\mathbf{Z}$.

Exercise 2:

1. Estimate $ \mathbf{E}^{do(X:=3)} [Y] $ from data.loaded.
2. We have regressed Y on X *and* $\mathbf{Z}$ --- thus the name ``adjusting for $\mathbf{Z}$". Do you see what goes wrong when we try to estimate the causal effect α by regressing Y on X without adjusting for $\mathbf{Z}$?

Solution 2:

In [ ]:

End of Solution 2

Whether or not a set is a valid adjustment set can be answered using the notion of d-separation (e.g., the set should block the path $X \leftarrow A \rightarrow K \rightarrow Y$, for example. Since d-separation is an important concept, we would like to revise it here.

In brief, given a DAG , the (disjoint) sets Q and R are d-separated by a set S if all paths between Q and R are blocked by S. A path $i_1, i_2, \ldots, i_m$ is blocked by S if there is a node i_k on the path ($1 < k < m$) such that one of the following conditions hold:

- ${i_k} \in {S}$ and \begin{align*} &{i_{k-1}} \rightarrow {i_k} \rightarrow {i_{k+1}}\\ \text{or }\;&{i_{k-1}} \leftarrow {i_k} \leftarrow {i_{k+1}}\\ \text{or }\;&{i_{k-1}} \leftarrow {i_k} \rightarrow {i_{k+1}} \end{align*}
- neither ${i_k}$ nor any of its descendants is in ${S}$ and $$ {i_{k-1}} \rightarrow {i_k} \leftarrow {i_{k+1}}. $$

The data that we have loaded come from an SCM. One can show that the distribution is Markov with respect to the corresponding graph. This means that d-separation implies conditional independence. We will try to verify this with a few examples.

Exercise 3:

Write down two d-separation statements that hold and two that do not hold. Test for conditional and unconditional independence using the functions `CondIndTest` from package `CondIndTests` and `dhsic.test` from package `dHSIC` , respectively. Use the data in data.loaded; you can access variables using `data.loaded[,"X"]` , for example. Do you find the correct conditional independences? Note that conditional independence testing is a difficult statistical problem, especially if the conditioning set is large.

Solution 3:

In [ ]:

```
```

End of Solution 3

If we are given the full SCM (this will often not be the case), there is a different strategy for computing causal effects. We will use this to validate our estimate from Exercise 2. In fact, the data that we have provided above, were generated by the following SCM: \begin{align*} C &:= N_C\\ A &:= 0.8 N_A\\ K &:= A + 0.1 N_K\\ X &:= C - 2 A + 0.2 N_X\\ F &:= 3 X + 0.8 N_F\\ D &:= -2 X + 0.5 N_D\\ G &:= D + 0.5 N_G\\ Y &:= 2 K - D + 0.2 N_Y\\ H &:= 0.5 Y + 0.1 N_H \end{align*} with all N's being jointly independent and having a standard normal distribution.

In [ ]:

```
set.seed(1); n <- 200
C <- rnorm(n)
A <- 0.8*rnorm(n)
K <- A + 0.1*rnorm(n)
X <- C - 2*A + 0.2*rnorm(n)
F <- 3*X + 0.8*rnorm(n)
D <- -2*X + 0.5*rnorm(n)
G <- D + 0.5*rnorm(n)
Y <- 2*K - D + 0.2*rnorm(n)
H <- 0.5*Y + 0.1*rnorm(n)
data.check <- cbind(C, A, K, X, F, D, G, Y, H)
data.loaded <- data.check
save(data.loaded, file = "./ComputeInterventionsData1.Rdata")
```

Indeed, this yields the same data:

In [ ]:

```
head(data.loaded)
head(data.check)
```

Exercise 4:

1. Generate data from the intervened SCM $do(X:=3)$ and compare your findings with the estimate computed above.
2. Look at the graph shown above and at the coefficients of the SCM. Can you guess the correct value for $\mathbf{E}^{do(X:=3)}[Y]$?

Solution 4:

In [ ]:

End of Solution 4

In [ ]: