InvariantCausalPrediction

June 23, 2019

1 Invariant Causal Prediction

by Jonas Peters, Niklas Pfister and Rune Christiansen, 18.06.2019

This notebook aims to give you a basic understanding of invariant causal prediction for causal inference.

The method's goal is as follows: Suppose we are given data $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ from a target variable Y and a vector of d predictors $\mathbf{X}$. We are then trying to determine the causal parents $\text{pa}(Y) \subseteq \{1, \dots, d\}$ of Y. The inference will be based on heterogeneity in the data (e.g., the data come from different interventional settings).

```
In [19]: library(InvariantCausalPrediction)
         library(seqICP)
```

1.1 Environment based approach

We first start with a fundamental observation that we will exploit later.

Assume the $d+1$ dimensonal vectors $\mathbf{Z}_i = (Z_i^0, Z_i^1, \dots, Z_i^d)$ for $i = 1, \dots, n$ are independent observations generated by (potentially) different interventional settings of the same linear structural causal model (SCM) such that the induced graphs are directed and acyclic (i.e., DAGs). Assume further that none of the interventions occurs directly on the variable Z^0. Then, for $Y := Z^0$ and $\mathbf{X} := (Z^1, \dots, Z^d)$ we have following invariance: There exists $\beta \in (\mathbb{R} \setminus \{0\})^{|\text{pa}(Y)|}$ such that for all $i \in \{1, \dots, n\}$ it holds that

$$Y_i = \mu + X_i^{\text{pa}(Y)}\beta + \epsilon_i \text{ and } \epsilon_i \perp\!\!\!\perp X_i^{\text{pa}(Y)}, \tag{1}$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. noise variables.

1.1.1 Exercise 1

Generate one sample from a distribution from the linear SCM

$$\mathcal{S} : \begin{cases} X_i = \epsilon_i^1 \\ Y_i = 1.5 \cdot X_i + \epsilon_i^2 \,, \end{cases} \tag{1}$$

and a second sample from the same SCM under a shift intervention on X. Plot both samples in the same (X,Y)-scatterplot using different colors. Does the conditional distribution of $Y|X$ remain invariant, i.e., it is the same in both samples? What about the distribution of Y?

1.1.2 Solution 1

```
In [20]: # Generate n=1000 observations from the observational distribution,
         # and store observations in vectors called "Xa" and "Ya"
         ######################
         # fill in
         ######################

         # Generate n=1000 observations from the interventional distribution,
         # and store observations in vectors called "Xb" and "Yb"
         ######################
         # fill in
         ######################

         if(exists("Xa") & exists("Xb") & exists("Ya") & exists("Yb")){
             # Plot both samples
             red <- rgb(1,0,0,alpha=0.4)
             blue <- rgb(0,0,1,alpha=0.4)
             # Y vs X1
             plot(Xa,Ya,pch=16,col=blue,xlim=range(c(Xa,Xb)),ylim=range(c(Ya,Yb)),xlab="X",ylab
             points(Xb,Yb,pch=17,col=red)
             legend("topright",c("observational","interventional"),pch=c(16,17),col=c(blue,red
         }
```

1.1.3 End Solution 1

We now assume that we are given the data and try to infer $\mathrm{pa}(Y)$. The method of invariant causal prediction exploits the invariance (1) from above. It goes over all sets of potential parents $\mathrm{pa}(Y)$ and finds all sets for which this invariance is satisfied.

To get a better understanding of how exactly invariant causal prediction performs this search, we consider the following toy data set.

```
In [21]: load(file = "./InvariantCausalPredictionData1.RData")  # load data
```

We have now loaded a sample consisting of the variables Y, X^1, X^2 and X^3. The variables correspond to the columns of the matrix data and the rows correspond to independent observations from an underlying SCM. The first 140 rows are sampled from an observational distribution, while the remaining 80 rows come from an interventional setting for which it is known that none of the interventions occured directly on Y. In the following two exercises we will determine the parents of Y using invariant causal prediction. First, we do this maually, and later we will make use of some functions already implemented in R.

1.1.4 Exercise 2

Perform a regression of Y on all possible sets of predictors (i.e. {X1}, {X2}, {X3}, {X1, X2}, {X1, X3}, {X2, X3}, {X1, X2, X3}). For each of the 7 regressions plot the residuals vs the fitted values (this is called a Tukey-Anscombe plot). In each figure, plot the data points from the first environment in "blue" and the points from the second environment in "red". Determine whether the corresponding conditional remains invariant across the two environments. Moreover, check whether

the distribution of Y itself remains invariant. What is the parent set? Hint: Think about which sets are definetly *not* the correct parent sets.

1.1.5 Solution 2

In [22]:
```
# extract response and predictors
Y <- data[,1]
Xmat  <- data[,2:4]

# define the potential parent sets
S <- list( c(1), c(2), c(3), c(1,2), c(1,3), c(2,3), c(1,2,3))

# perform regression for each set in S
resid <- fitted <- vector("list", length(S))
for(i in 1:length(S)){
  # regress Y linearly on the i'th set S (e.g. using lm.fit)
  # store the residuals in resid[[i]]
  # and the fitted values in fitted[[i]]
  ##########################
  ## fill in
  ##########################
}

filledInResid <- all(unlist(lapply(resid, length)) == length(Y))
filledInFitted <- all(unlist(lapply(fitted, length)) == length(Y))

if(filledInResid & filledInFitted){
    # plot the resulting
    env <- c(rep(0,140),rep(1,80))
    par(mfrow=c(2,2))
    red <- rgb(1,0,0,alpha=0.4)
    blue <- rgb(0,0,1,alpha=0.4)
    names <- c("X1", "X2", "X3", "X1, X2", "X1, X3", "X2, X3", "X1, X2, X3")
    # plot Y vs index (empty set)
    plot((1:length(Y))[env==0], Y[env==0], pch=16, col=blue, xlim=c(0,220), ylim=rang
    points((1:length(Y))[env==1], Y[env==1], pch=17, col=red)
    legend("topleft",c("observational","interventional"),pch=c(16,17),col=c(blue,red)
    # all remaining potential sets
    for(i in 1:length(S)){
        plot(fitted[[i]][env==0], resid[[i]][env==0], pch=16, col=blue, xlim=range(fi
        points(fitted[[i]][env==1], resid[[i]][env==1], pch=17, col=red)
        legend("topleft",c("observational","interventional"),pch=c(16,17),col=c(blue,
    }
}
```

1.1.6 End of Solution 2

1.1.7 Exercise 3

For the same data set apply the invariant causal prediction function ICP form the package InvariantCausalPrediction to determine the parent set. Hint: You will need to define a vector ExpInd which has the same length as the number of observations and indicates from which environment each observations comes (e.g. 0 for observational data and 1 for interventional data).

1.1.8 Solution 3

1.1.9 End of Solution 3

1.2 Extension to an environment-free approach

In the above exercises we knew which observations corresponded to the observational and which to the interventional setting. In this section we want to show that we can still apply a similar methodology even if this environment information is not known. All we need is a sequential ordering of the data. For example, the data could be grouped together for each environment or the interventions could change continuously across time. We illustrate this using the following toy example.

```
In [23]: load(file = "./InvariantCausalPredictionData2.RData")  # load data2
```

The matrix data2 contains the three variables Y, X^1 and X^2 as columns and each row corresponds to an independent observations from the same SCM under smoothly changing interventions. To be more precise, the interventions correspond to smooth shifts in the variance of the noise.

1.2.1 Exercise 4

Use the invariant causal prediction function for sequential data seqICP from the package seqICP to find an estimate of the parent set for the variable Y. Set the parameter test to "smooth.variance", this leads the seqICP to performs a hypothesis test tuned against alternatives that result from smooth variance interventions.

1.2.2 Solution 4

1.2.3 End of Solution 4

1.3 References

[1] Peters, J., P. Bühlmann, and N. Meinshausen (2016). *Causal inference using invariant prediction: identification and confidence intervals*. Journal of the Royal Statistical Society, Series B (with discussion) 78 (5), 947–1012.

[2] Pfister, N., P. Bühlmann and J. Peters (2018). *Invariant Causal Prediction for Sequential Data*. Journal of the American Statistical Association (accepted), ArXiv e-prints (1706.08058).