Predicting the effect of interventions using invariance principles for nonlinear models

Christina Heinze-Deml
Seminar for Statistics, ETH Zürich
heinzedeml@stat.math.ethz.ch

Jonas Peters
University of Copenhagen
jonas.peters@math.ku.dk

Nicolai Meinshausen
Seminar for Statistics, ETH Zürich
meinshausen@stat.math.ethz.ch

Abstract

An important problem in many domains is to predict how a system will respond to interventions. This task is inherently linked to estimating the system's underlying causal structure. To this end, Invariant Causal Prediction (ICP) [10] has been proposed which learns a causal model exploiting the invariance of causal relations in different environments. When considering linear models, the implementation of ICP is relatively straightforward; the nonlinear case, however, is more challenging due to the difficulty of performing nonparametric tests for conditional independence. In this work, we present one method for nonlinear Invariant Causal Prediction. As an example, we consider fertility rate modeling which is central to world population projections. We explore predicting the effect of hypothetical interventions using the accepted models from nonlinear ICP.

1 Introduction

Developing countries have a significantly higher fertility rate compared to Western countries. The fertility rate can be predicted well from covariates such as 'infant mortality rate' or 'GDP per capita'. Classical prediction models, however, do not answer whether an active intervention on some of the covariates leads to a change in the fertility rate. This can only be answered by exploiting causal knowledge of the system. Traditionally, in statistics the methods for establishing causal relations rely on carefully designed randomized studies. Often, however, such experiments cannot be performed. For instance, factors like 'infant mortality rate' are highly complex and cannot be manipulated in isolation. We may still be interested in the effect of a policy that aims at reducing the infant mortality rate but this policy cannot be randomly assigned to different groups of people within a country. At the hand of the example from fertility rate modeling, we shall explore how to exploit the invariance of causal models for causal discovery in the nonlinear case.

1.1 Notation

Assume an underlying structural equation model (SEM) [e.g. 9]

$$
\begin{aligned}
Z_1 &\leftarrow g_1(Z_{\mathrm{pa}_1}) + \eta_1, \\
Z_2 &\leftarrow g_2(Z_{\mathrm{pa}_2}) + \eta_2, \\
&\vdots \\
Z_q &\leftarrow g_q(Z_{\mathrm{pa}_q}) + \eta_q,
\end{aligned}
$$

for which the functions g_k, $k = 1, \ldots, q$, as well as the parents $\mathrm{pa}_k \subseteq \{1, \ldots, q\} \setminus \{k\}$ of each variable are unknown. Here, we have used the notation $Z_S = (Z_{i_1}, \ldots, Z_{i_s})$ for any set $S = \{i_1, \ldots, i_s\} \subseteq \{1, \ldots, q\}$. We assume the corresponding directed graph to be acyclic. We further require the noise variables $\eta_1, \ldots, \eta_q$ to be jointly independent and to have zero mean.

Instead of trying to infer the whole graph [e.g. 9, 13, 3, 11, 6, 5] we are here interested in settings where there is a target variable Y of special interest. W.l.o.g., we write $Y = Z_1$ taking values in $\mathcal{Y}$ and denote all the other covariates by $X := Z_{\{2,\ldots,q\}}$. We further write $S^* := \mathrm{pa}_1$ for the parents of Y and $\varepsilon := \eta_1$. Thus, the structural equation for Y has the form

$$Y \leftarrow f(X) + \varepsilon, \tag{1}$$

where $f : \mathbb{R}^p \to \mathcal{Y}$ with $p = q - 1$. We let $\mathcal{F}$ be the function class of f and let $\mathcal{F}_S$ be the subclass of functions that depend only on the set $S \subseteq \{1, \ldots, p\}$ of variables. We thus have $f \in \mathcal{F}_{S^*}$.[1] The goal is to infer both the parental set S^* and confidence bands for the function f.

1.2 Invariance approach for causal discovery

[10] proposed an invariance approach in the context of linear models. We describe the approach here in a notationally slightly different way that will simplify statements and results in the nonlinear case and allow for more general applications.

Definition 1 (Environmental variables) *Let X_E with $E \subseteq \{1, \ldots, p\}$ be a subset of the predictor variables X, namely the so-called environmental variables X_E, such that we know or assume that X_E are not descendants of Y in the causal DAG of (X, Y).*

In [10], the environmental variables were given and non-random. Note that the definition above leaves open the possibility that there is a direct causal connection between one of the variables in X_E and Y, in contrast to, say, instrumental variable approaches.

Definition 2 (Strict environmental variables) *We say the environmental variables X_E are strict environmental variables if there is no direct edge from X_E to the outcome Y of interest.*

Example (Fertility data) In this work, we analyze a data set provided by [15]. Here, Y, X and X_E correspond to the following quantities:

(i) $Y \in \mathbb{R}$ is the total fertility rate in a country in a given year,

(ii) $X \in \mathbb{R}^p$ with $p = 9$ are potential causal predictor variables for TFR:
   – IMR – infant mortality rate
   – Q5 – under-five mortality rate
   – Education expenditure (% of GNI)
   – Exports of goods and services (% of GDP)
   – GDP per capita (constant 2005 US$)
   – GDP per capita growth (annual %)
   – Imports of goods and services (% of GDP)
   – Primary education (% female)
   – Urban population (% of total)

(iii) $X_E \in \mathbb{R}$ is the continent of the country, divided into the categories Africa, Asia, Europe, North and South America and Oceania. If viewed as a random variable (which one can argue about), the assumption is that the continent is not a descendant of the fertility rate, which seems plausible. For a strict environmental variable, the additional assumption is that the TFR in a country is only indirectly (that is via one of the other variables) influenced by which continent it is situated on (cf. Figure 1).

The basic yet central insight underlying the invariance approach is the fact that for the true causal parental set S^* and environmental variables X_E we have the following conditional independence relation

$$Y \perp\!\!\!\perp X_E \mid X_{S^*}. \tag{2}$$

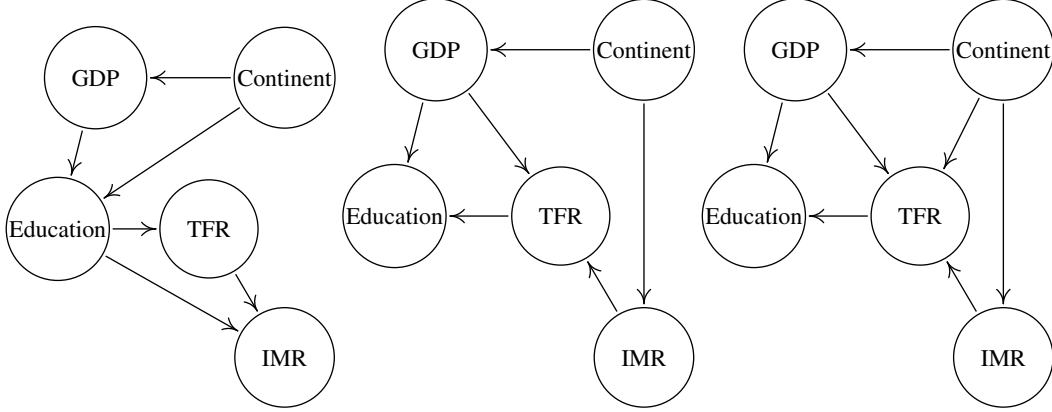[1]In slight abuse of notation, we identify X_{S^*} with the variables Z_{S^*}.

Figure 1: Three examples of a causal DAG with target total fertility rate (TFR) and four potential causal predictor variables. We would like to infer the parents of TFR in the graph. The environment variable X_E is the continent. If we assume that the continent has no direct causal influence on TFR (as in the two DAGs on the left) we say that the environmental variable is strict.

In case of an SEM, for example, this follows directly from the local Markov condition. In general, some components in E and S^* can be identical as $E \cap S^* \neq \emptyset$. However, in this work we assume strict environmental variables, which implies $E \cap S^* = \emptyset$. The goal is to find S^* by exploiting the above relation. Suppose we have a test for the null hypothesis

$$H_{0,S}: \qquad Y \perp\!\!\!\perp X_E \mid X_S. \tag{3}$$

It was then proposed in [10] to define an estimate $\hat{S}$ for the parental set S^* by setting

$$\hat{S} := \bigcap_{S:H_{0,S} \text{ not rejected}} S. \tag{4}$$

Here, the intersection runs over all sets S, s.t. $E \cap S = \emptyset$. If the tests are conducted at level $\alpha \in (0,1)$, we have the coverage guarantee that

$$P(\hat{S} \subseteq S^*) \geq 1 - \alpha,$$

which follows directly from the fact that S^* is accepted with probability at least $1 - \alpha$ since H_{0,S^*} is true; see [10] for details. The conditional independence (3) was tested in [10] under the assumption of a linear model (and strict environmental variables).

1.3 Contribution

Our contributions are fourfold:

(i) **Conditional independence test.** We propose in Section 2.1 one possible nonlinear and non-parametric test for conditional independence of the type (3). There has been some progress towards nonparametric independence tests [1, 7, 2, 12, 14, 16]. However, in the general non-parametric case, no known test of conditional independence has (even asymptotically) a type I error rate less than the pre-specified significance level. This stresses the importance of empirical evaluation of conditional independence tests.

(ii) **Defining sets.** We discuss in Section 2.2 cases of poor identifiability of the causal parents. If there are highly correlated variables in the dataset, we might get an empty estimator $\hat{S} = \emptyset$ for the causal parents (or an $\hat{S}$ with low cardinality). We can, however, extract more information via defining sets comparable to a similar issue arising in multiple testing [4]. For example, if the sets $S = \{1,3\}$ and $S = \{2,3\}$ are accepted in (3), we have that $\hat{S} = \{3\}$ and, with high probability, X_3 is causal for the target Y. Yet we also know that either X_1 or X_2 has to be causal for the target variable.

(iii) **Confidence bands for causal effects.** Beyond identifying the causal parents, we can provide nonparametric or nonlinear confidence bands for the strength of the causal effects, as shown in Section 2.3.

3

(iv) **Prediction under interventions.** Using the accepted models from nonlinear ICP, we are able to forecast the average causal effect of external interventions. We will discuss this at hand of examples in Section 2.4.

2 Methodology

2.1 Conditional independence tests

The confidence set $\hat{S}$ for S^* can be constructed in the nonparametric setting analogous to the linear case by defining

$$\hat{S} := \bigcap_{S:H_{0,S} \text{ not rejected}} S \tag{5}$$

where

$$H_{0,S} : Y \perp\!\!\!\perp X_E \mid X_S. \tag{6}$$

Because of (2), the null hypothesis is true for the true set S^* of causal parents. If we can guarantee that the test of (6) has the correct type I error rate in the sense that

$$P\big(H_{0,S^*} \text{ is rejected at level } \alpha\big) \leq \alpha, \tag{7}$$

then we have as immediate consequence the desired statement

$$P\big(\hat{S} \subseteq S^*\big) \ \geq \ P\big(H_{0,S^*} \text{ accepted}\big) \ \geq \ 1 - \alpha.$$

If we assume a linear function f in the structural equation (1), then tests that can guarantee the level as in (7) are available [see 10]. However, in the general nonlinear and nonparametric case, it becomes more difficult to guarantee the type I error rate. In a nonlinear setting, where we know an appropriate basis expansion of f, we can of course revert back to the linear case. Apart from such special circumstances, we have to find tests that guarantee the type I error rate in (7) as closely as possible under a wide range of scenarios. In this work, we consider the following method:

Invariant conditional quantile prediction. Predict a $1 - \beta$ quantile of the conditional distribution of Y, given X, by pooling the data over all environments and using a Quantile Regression Forest [8]. Then test whether the exceedance of the conditional quantiles is independent of the environment variable, using Fisher's exact test across all discrete environments. This yields a p-value $p(\beta)$ for the $1 - \beta$ quantile. Repeat for a number of quantiles and aggregate the resulting individual p-values $p(\beta)$ by Bonferroni correction. For instance, predict the $1 - \beta$ quantiles for $\beta \in \{0.1, 0.5, 0.9\}$ and compute the overall p-value as $p = 3 \cdot \min_{\beta \in \{0.1, 0.5, 0.9\}} p(\beta)$.

Invariant conditional quantile prediction does not require the noise variable in (1) to be additive. However, since additive noise is used in Sections 2.3 and 2.4, we have written the structural equations in an additive form and shall assume additive noise throughout. In an extended version of this work, we analyze the type I error rate of the proposed test empirically. We further propose and empirically evaluate various other methods to test the null hypothesis of conditional independence in a nonlinear setting. One of the inherent difficulties with these tests is that the bias in the estimation of the optimal function can potentially lead to a more frequent rejection of a true null hypothesis than the nominal level suggests.

Example (Fertility data) The following sets were accepted at the level $\alpha = 0.1$ when using nonlinear ICP with invariant conditional quantile prediction:

$S_1 = \{\text{Q5}\}$
$S_2 = \{\text{IMR, Imports of goods and services, Urban pop. (\% of total)}\}$
$S_3 = \{\text{IMR, Education expenditure (\% of GNI), Exports of goods and services, GDP per capita}\}$

As the intersection of $S_1, \ldots, S_3$ is empty, we have $\hat{S} = \emptyset$. This motivates the concept of defining sets.

4

2.2 Defining sets

It is often impossible to distinguish between highly correlated variables. For example, infant mortality IMR and under-five mortality Q5 are highly correlated in the data and can often be substituted for each other. We thus accept sets that contain either of these variables. When taking the intersection as in (5), this leads to exclusion of both variables in $\hat{S}$ and potentially to an altogether empty set $\hat{S}$. We can instead ask for the defining sets [4], where a defining set $\hat{D} \subseteq \{1, \ldots, p\}$ has the properties

(i) $S \cap \hat{D} \neq \emptyset$ for all S such that $H_{0,S}$ is accepted.

(ii) there exists no strictly smaller set D' with $D' \subset \hat{D}$ for which property (i) is true.

Given a defining set $\hat{D}$, we thus know that

$$P(S^* \cap \hat{D} = \emptyset) \leq P(H_{0,S^*} \text{ rejected}) \leq \alpha.$$

In other words, a) at least one of the variables in the defining set $\hat{D}$ is a parent of the target, and b) the data do not allow to resolve it at a less granular level.

Example (Fertility data)   We obtain seven defining sets:

$\hat{D}_1 = \{\text{Q5, IMR}\}$
$\hat{D}_2 = \{\text{Q5, Education expenditure (\% of GNI), Imports of goods and services}\}$
$\hat{D}_3 = \{\text{Q5, Education expenditure (\% of GNI), Urban pop. (\% of total)}\}$
$\hat{D}_4 = \{\text{Q5, Exports of goods and services, Imports of goods and services}\}$
$\hat{D}_5 = \{\text{Q5, Exports of goods and services, Urban pop. (\% of total)}\}$
$\hat{D}_6 = \{\text{Q5, GDP per capita, Imports of goods and services}\}$
$\hat{D}_7 = \{\text{Q5, GDP per capita, Urban pop. (\% of total)}\}$

2.3 Confidence bands

For a given set S, we can in general construct a $(1 - \alpha)$-confidence band $\hat{\mathcal{F}}_S$ for the regression function when predicting Y with the variables X_S. Note that f is the regression function when regressing Y on the true set of causal variables X_{S^*} and hence, with probability $1 - \alpha$, we have

$$P(f \in \hat{\mathcal{F}}_{S^*}) \geq 1 - \alpha.$$

Furthermore, from Section 1.2 we know that H_{0,S^*} is accepted with probability $1 - \alpha$. We can hence construct a confidence band for the causal effects as

$$\hat{\mathcal{F}} := \bigcup_{S:H_{0,S}\text{not rejected}} \hat{\mathcal{F}}_S. \tag{8}$$

Using a Bonferroni correction, we have the guarantee that

$$P\big(f \in \hat{\mathcal{F}}\big) \ \geq \ 1 - 2\alpha,$$

where the coverage guarantee is point-wise or uniform, depending on the coverage guarantee of the underlying estimators $\hat{\mathcal{F}}_S$ for all given $S \subseteq \{1, \ldots, p\}$.

2.4 Predicting the effect of interventions

The confidence bands $\hat{\mathcal{F}}$ themselves can be difficult to interpret. Interpretability can be guided by looking at the average causal effect in the sense that we compare the expected response at two points $x, \tilde{x} \in \mathbb{R}^p$:

$$\Delta(x, \tilde{x}) \ := \ E\big(Y\big|\text{do}(X = x)\big) - E\big(Y\big|\text{do}(X = \tilde{x})\big). \tag{9}$$

For the fertility data, this would involve a hypothetical scenario where, for the first term on the r.h.s. in (9), we fix the variables to be equal to x for a country and, for the second term, we set the variables to $\tilde{x}$, which might differ from x just in one or a few coordinates. Eq. (9) then compares the average
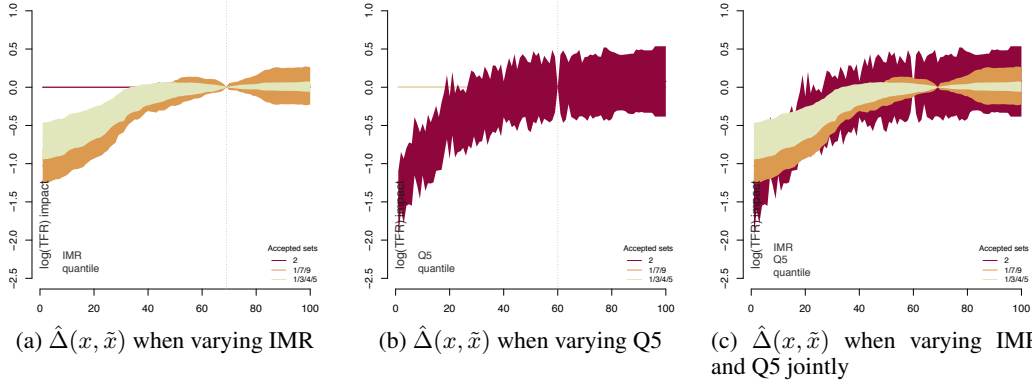
(a) $\hat{\Delta}(x, \tilde{x})$ when varying IMR   (b) $\hat{\Delta}(x, \tilde{x})$ when varying Q5   (c) $\hat{\Delta}(x, \tilde{x})$ when varying IMR and Q5 jointly

Figure 2: Data for Nigeria in 1993: The union of the confidence bands $\hat{\mathcal{F}}_S$, denoted by $\hat{\mathcal{F}}$, bounds the average causal effect of varying the variables in the defining set $\hat{D}_1 = \{\text{IMR, Q5}\}$ on the target $\log(\text{TFR})$. IMR and Q5 have been varied individually, see panels (a) and (b), as well as jointly, see panel (c), over their respective quantiles. The chosen significance level is $\alpha = 0.1$, i.e. $P\big(\Delta(x, \tilde{x}) \in \hat{\Delta}(x, \tilde{x})\big) \geq 0.8$.

expected fertility between these two scenarios. Note that the expected response under a do-operation is just a function of the causal variables $S^* \subseteq \{1, \ldots, p\}$,

$$E\big(Y \big| \text{do}(X = x)\big) = E\big(Y \big| \text{do}(X_{S^*} = x_{S^*})\big).$$

In the absence of hidden variables, we even have that the latter is equal to

$$E\big(Y \big| \text{do}(X_{S^*} = x_{S^*})\big) = E\big(Y \big| X_{S^*} = x_{S^*}\big),$$

that is it does not matter whether we set the causal variables to a specific value x_{S^*} or whether they were observed in this state.

Once we have a confidence band as defined in (8), we can bound the average causal effect (9) by the interval

$$\hat{\Delta}(x, \tilde{x}) := \Big[\inf_{f \in \hat{\mathcal{F}}} \{f(x) - f(\tilde{x})\}, \ \sup_{f \in \hat{\mathcal{F}}} \{f(x) - f(\tilde{x})\}\Big],$$

with the immediate guarantee that

$$P\Big(\Delta(x, \tilde{x}) \in \hat{\Delta}(x, \tilde{x})\Big) \geq 1 - 2\alpha.$$

Example (Fertility data) The confidence bands $\hat{\mathcal{F}}$, required for the computation of $\hat{\Delta}(x, \tilde{x})$, are obtained by a time series bootstrap as the fertility data contain temporal as well as spatial dependencies. We use a level of $\alpha = 0.1$ for the hypothesis tests as well as for the confidence intervals, so that $P\big(\Delta(x, \tilde{x}) \in \hat{\Delta}(x, \tilde{x})\big) \geq 0.8$. In the examples below, we set x to an observed data point and vary only $\tilde{x}$.

In the first example, we consider the observed covariates for Nigeria in 1993 as x. The point of comparison $\tilde{x}$ is set equal to x, except for the variables in the defining set $\hat{D}_1 = \{\text{IMR, Q5}\}$. In Figures 2(a) and (b), these are varied individually over their respective quantiles. The overall confidence interval $\hat{\mathcal{F}}$ consists of the union of the shown confidence intervals $\hat{\mathcal{F}}_S$. If $x = \tilde{x}$ (depicted by the vertical lines), the average causal effect is zero, of course. In neither of the two scenarios shown in Figures 2(a) and (b), we observe consistent effects different from zero as some of the accepted models do not contain IMR resp. Q5. However, when varying the variables $\hat{D}_1 = \{\text{IMR, Q5}\}$ jointly (see Figure 2(c)), we see that all accepted models predict an increase in expected $\log(\text{TFR})$ as IMR and Q5 increase.

In the second example, we compare the expected fertility rate between countries where all covariates are set to the value x, which is here chosen to be equal to the observed values of all African countries in 2013. The expected value under this value x of covariates is compared to the scenario where we take as $\tilde{x}$ the same value but set the values of the variables IMR and Q5 to the respective European
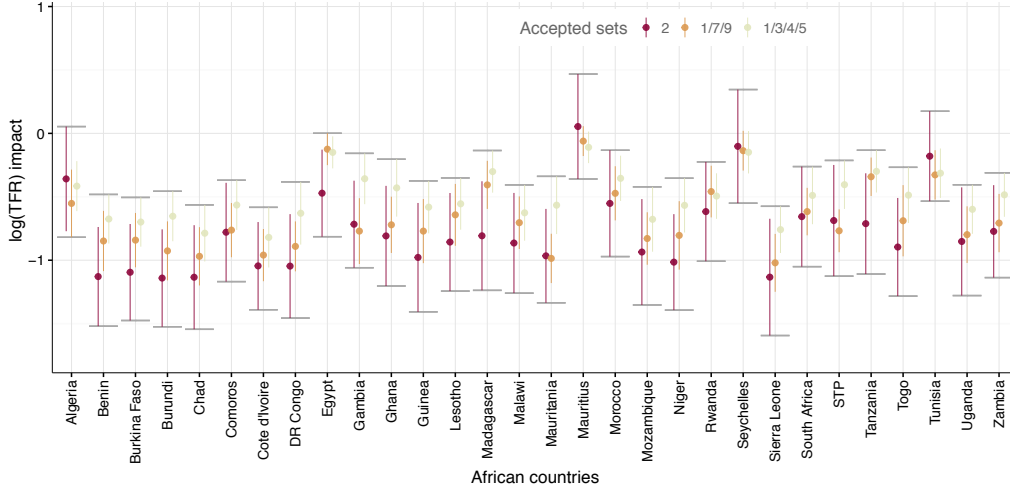
(a) $\hat{\Delta}(x, \tilde{x})$ estimated using nonlinear ICP with invariant conditional quantile prediction
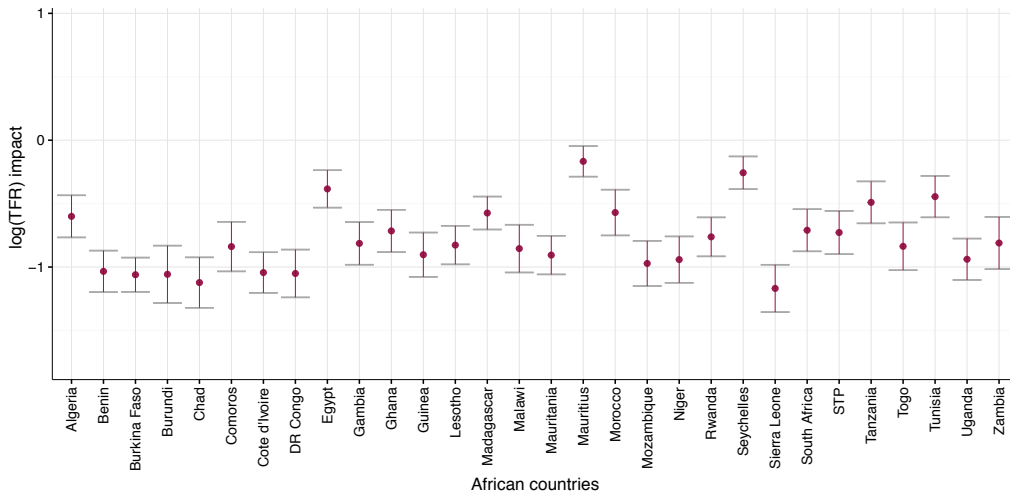
(b) Random forest regression model

Figure 3: (a) Bounds for the average causal effect of setting the variables IMR and Q5 in the African countries in 2013 to European levels (with $P\big(\Delta(x, \tilde{x}) \in \hat{\Delta}(x, \tilde{x})\big) \geq 0.8$). (b) Random forest regression model using all covariates as input (with 80%-confidence intervals).

averages. The union of intervals in Figure 3(a) (depicted by the horizontal line segments) correspond to $\hat{\Delta}(x, \tilde{x})$ for each country under nonlinear ICP with invariant conditional quantile prediction. The accepted models make largely coherent predictions for the effect associated with this comparison. For most countries, the difference is negative, meaning that the average expected fertility declines if the child mortality rate in a country decreases to European levels. The countries where $\hat{\Delta}(x, \tilde{x})$ contains 0 typically have a child mortality rate that is close to European levels, meaning that there is no substantial difference between the two points $x, \tilde{x}$ of comparison.

For comparison, in Figure 3(b), we show the equivalent computation as in Figure 3(a) when *all* covariates are *assumed* to have a causal effect on the target and a random forest is used for estimation. While the resulting regression bootstrap confidence intervals often overlap with $\hat{\Delta}(x, \tilde{x})$, they are typically much smaller. This implies that if the regression model containing all covariates was – wrongly – used as a surrogate for the causal model, the uncertainty of the prediction would be underestimated. Furthermore, such an approach ignoring the causal structure can lead to a significant bias in the prediction of causal effects when we consider interventions on descendants of the target variable, for example.

3 Discussion and future work

In this work, we have proposed invariant conditional quantile prediction, a nonparametric test for conditional independence of the type (3). We have used this test for nonlinear ICP to model how several interventions may affect the total fertility rate of a country. In particular, we provided bounds on the average causal effect. For some countries, the effect can be bounded away from zero when interventions are performed simultaneously on infant mortality rate and under-five mortality rate.

In an extended version of this work, we aim to propose and empirically evaluate alternative tests for assessing the invariance of conditional distributions for nonlinear models. Additionally, we would like to discuss the case of a direct effect of environmental variables on the outcome of interest Y. Under certain assumptions the causal parents of Y can then still be identifiable. As this allows for a direct causal connection between one of the variables in X_E and Y, this can be a key advantage over standard instrumental variable approaches.

References

[1] W. Bergsma and A. Dassios. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20:1006–1028, 2014.

[2] J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32:485–498, 1961.

[3] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[4] J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, pages 584–597, 2011.

[5] A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B*, 77:291–318, 2015.

[6] D. Heckerman. A Bayesian approach to causal discovery. Technical report, Microsoft Research (MSR-TR-97-05), 1997.

[7] W. Hoeffding. A non-parametric test of independence. *Ann. Math. Statist.*, 19:546–557, 12 1948.

[8] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[9] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

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

[11] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[12] A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.

[13] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, USA, 2nd edition, 2000.

[14] G. J. Szkely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007.

[15] United Nations. World population prospects: The 2012 revision. *Population Division, Department of Economic and Social Affairs, United Nations, New York*, 2013.

[16] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.