Invariance, Causality and Robustness

2018 Neyman Lecture [*]

Peter Bühlmann [†]
Seminar for Statistics, ETH Zürich

December 21, 2018

Abstract

We discuss recent work for causal inference and predictive robustness in a unifying way. The key idea relies on a notion of probabilistic invariance or stability: it opens up new insights for formulating causality as a certain risk minimization problem with a corresponding notion of robustness. The invariance itself can be estimated from general heterogeneous or perturbation data which frequently occur with nowadays data collection. The novel methodology is potentially useful in many applications, offering more robustness and better "causal-oriented" interpretation than machine learning or estimation in standard regression or classification frameworks.

Keywords: Anchor regression, Causal regularization, Distributional robustness, Heterogeneous data, Instrumental variables regression, Interventional data, Random Forests, Variable importance.

1 Introduction

Understanding the causal relationships in a system or application of interest is perhaps the most desirable goal in terms of understanding and interpretability. There is a rich history of developments from various disciplines, dating back to ancient times: "Felix, qui potuit rerum cognoscere causas" – Fortunate who was able to know the causes of things (Georgics, Virgil, 29 BC). One might think that for pure prediction tasks, without any ambition of interpretability, knowing the causes or the causal structure is not important. We will explain here how these problems are related and as a consequence: (i) one can obtain "better" predictions when incorporating causal aspects and (ii) one can infer causal structure from a certain predictive perspective.

Inferring causal structure and effects from data is a rapidly growing area. When having access to data from fully randomized studies, Jerzy Neyman made a pioneering contribution using a potential outcome model (Splawa-Neyman, 1990).

Randomized studies serve as the gold standard and the corresponding inference of causal effects can be viewed as "confirmatory" due to the fact that the underlying model assumptions are not substantially more restrictive than for say a standard regression type problem, see for example Dawid (2000); Pearl (2009); Hernán and Robins (2010); Rubin and Imbens (2015); VanderWeele

[*] This article is covering material of the 2018 Neyman Lecture but also describes new developments in Section 5.

[†] This project has received funding from the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG).

Figure 1: Jerzy Neyman (1894-1981). Besides other pioneering work, he has also made fundamental early contributions to causality in 1923, in terms of mathematical formulation with the potential outcome model (Splawa-Neyman, 1990). Left: taken from `http://www.learn-math.info/history/photos/Neyman_3.jpeg`. Right: taken from `https://errorstatistics.com/2017/04/16/a-spanos-jerzy-neyman-and-his-enduring-legacy-3/` by A. Spanos. The photograph is hanging on the wall in the coffee room of the Department of Statistics at UC Berkeley.

(2015). Often though, the data at hand does not come from a (fully) randomized study: the question is now whether one can still infer causal effects and under what kind of assumptions this is possible. A range of different approaches have been suggested, see for example Greenland et al. (1999); Robins et al. (2000); Spirtes et al. (2000); Richardson et al. (2002); Hernán and Robins (2006); Tchetgen and VanderWeele (2012); Chickering (2002); Kalisch and Bühlmann (2007); Maathuis et al. (2009); Hauser and Bühlmann (2015), exhibiting different degrees of "confirmatory" nature for inferring causal effects. Since causal inference is very ambitious, these techniques should be thought as "geared towards causality" but not necessarily able to infer the underlying true causal effects. Still, the point is that they do something "more intelligent towards causality" than an analysis based on a standard potentially nonlinear regression or classification framework. We believe that this is an important area in statistics and machine learning: in particular, these techniques often have a more "causal-type" and thus more interesting interpretation than standard machine learning methods and therefore, this topic is important in the advent of "interpretable machine learning".

1.1 A framework based on invariance properties

We will focus here on a particular framework with corresponding methods which are "geared towards" causal solutions: with stronger assumptions (but less strong than for some competitor methods) they infer causal effects while under more relaxed and perhaps more realistic assumptions, they are still providing solutions for a "diluted form of causality"[1] which are often more meaningful than what is provided by regression or classification techniques. This can be made mathematically more rigorous, in terms of a novel form of robustness.

The construction of methods ~~is~~ relies on exploiting invariance from heterogeneous data. The heterogeneity can be unspecific perturbations and in this sense, the current work adds to the still yet quite small literature on statistics for perturbation data.

[1] I am grateful to Ed George who suggested this term.

1.2 Our contribution

The first part of the manuscript is a review of our own work in Peters et al. (2016) and Rothenhäusler et al. (2018) but putting the contributions into a broader perspective. We also add some novel methodology on nonlinear anchor regression and present some corresponding illustrations in Section 5.

2 Predicting potential outcomes, heterogeneity and worst case risk optimization

Predicting potential outcomes is a relevant problem in many application areas. Causality deals with a quantitative answer (a prediction) to a "What if I do question" or a "What if I perturb question".

2.1 Two examples for prediction of potential outcomes

The first example is from genomics (Stekhoven et al., 2012). The response variable of interest is the flowering time of the *Arabidopsis thaliana* plant (the time it takes until the plant is flowering) and the covariates are gene expressions from 21'326 genes, that is from a large part of the genome. The problem is to predict the flowering time of the plant when making single gene interventions, that is, when single genes are "knocked out". The data is from the observational state of the system only without any interventions. Therefore, this is a problem of predicting a *potential* outcome which has never been observed in the data. Even if one fails to infer the true underlying causal effects when making interventions, our viewpoint is that a prediction being better than from a state-to-the art regression method is still very useful for e.g. prioritizing experiments which can be done subsequently in a biology lab, see for example Maathuis et al. (2010) and the Editorial (2010).

The second example is about predicting behavior of individuals when being treated by advertisement campaigns. Such an advertisement could happen on social media for political campaigns or various commercial products. Consider the latter, namely an advertisement for commercial products on social media. The response of interest is how deep an individual user clicks on the advertisement and the subsequent web-pages, the covariates are attributes of the user. The task is to predict the response if one would intervene and show to a certain user "X" a certain advertisement "A": but there is no data for user "X" or similar users as "X" being exposed (or treated) with advertisement "A". Thus, it is a problem of predicting a *potential* outcome which has never been observed in the data. As mentioned in the genomic example above, even if we cannot infer the underlying true causal effect of treatment with an advertisement, it is still informative and valuable to come up with a good prediction for the response under an intervention which we have never seen in the data. See for example Bottou et al. (2013) or also Brodersen et al. (2015).

2.2 The heterogeneous setting with different environments

We consider data from different *observed (known)* environments, and we sometimes refer to them also as experimental settings or sub-populations or perturbations:

$$(\mathbf{Y}^e, \mathbf{X}^e), e \in \mathcal{E} \tag{1}$$

with $n_e \times 1$ response vectors $\mathbf{Y}^e$, $n_e \times p$ covariate design matrices $\mathbf{X}^e$ and e denotes an environment from the space $\mathcal{E}$ of observed environments. Here, n_e denotes the sample size in environment e. We assume that the n_e samples in environment e are i.i.d. realizations of a univariate random variable Y^e and a p-dimensional random vector X^e.

Examples. As a first example, consider data from 10 different countries where we know the correspondence of each data point to one of the 10 countries. Then the space of observed (known) environments can be encoded by the labels from $\mathcal{E} = \{1, 2, \ldots, 10\}$. As a second example, consider economical data which is collected over time. Different environments or sub-populations then correspond to different blocks of consecutive time points. Assuming that these blocks are known and given, the space $\mathcal{E}$ then contains the labels for these different blocks of sub-populations.

Heterogeneity can also occur outside the observed data. Thus, we consider a space of unobserved environments

$$\mathcal{F} \supset \mathcal{E} \tag{2}$$

which is typically much larger than the space of observed environments $\mathcal{E}$.

Examples (cont.). In the examples from above, the space $\mathcal{F}$ could be: the 10 countries which are observed in the data and all other countries in the world; or the sub-populations of economical scenarios which we have observed in the data until today and all sub-populations of future scenarios which we have not seen in the data.

A main task is to make predictions for new unseen environments $e \in \mathcal{F}$ as discussed next.

2.3 A prediction problem and worst case risk optimization

We consider the following prediction problem.

Predict Y^e given X^e such that the prediction "works well" or is "robust" for all $e \in \mathcal{F}$ based on data from much fewer environments $e \in \mathcal{E}$.

Note that $\mathcal{F} \setminus \mathcal{E}$ is non-observed. The meaning of the aim above is that one is given in the future new covariates X^e from $e \in \mathcal{F} \setminus \mathcal{E}$ and the goal is to predict the corresponding Y^e. The terminology "works well" or is "robust" is understood here in the sense of performing well in worst-case scenarios. We note that the problem above is also related to transfer learning (Pratt, 1993; Pan and Yang, 2010; Rojas-Carulla et al., 2018).

In a linear model setting, this prediction task exhibits a relation to the following worst case L_2-risk optimization:

$$\text{argmin}_b \max_{e \in \mathcal{F}} \mathbb{E}[|Y^e - X^e b|^2]. \tag{3}$$

This problem has an interesting connection to causality. Before giving a more rigorous formulation, we describe the connection in a more loose sense, for the purpose of easier understanding. We consider the class $\mathcal{F}$ which includes all heterogeneities or perturbations e fulfilling two main assumptions:

ad-hoc condition 1: e does not act directly on Y^e.

ad-hoc condition 2: e does not change the mechanism between X^e and Y^e.

ad-hoc aim: ideally, e should change the distribution of X^e.

The ad-hoc conditions 1 and 2 are formulated precisely in assumption (B($\mathcal{F}$)) in Section 3.1. Regarding the ad-hoc aim: if there are many e which change the distribution of X^e, this introduces more observed heterogeneities in $\mathcal{E}$ which in term is favorable for better identification of causal effects.

Figure 2 is a graphical illustration of the ad-hoc conditions and aim above. For this purpose, we may think that the environments e are generated from a random variable E. We remark that there could be also hidden confounding variables between X and Y: more details are given in Section 4.

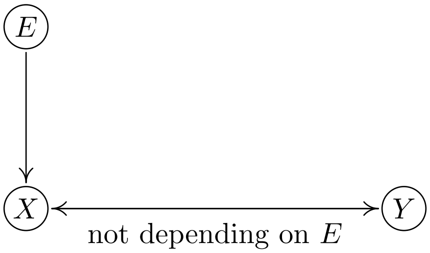


Figure 2: Graphical illustration of the ad-hoc conditions 1, 2 and the ad-hoc aim. There could be also hidden confounding variables, see Section 4.

An interesting connection to causality is then as follows:

$$\operatorname{argmin}_b \max_{e \in \mathcal{F}} \mathbb{E}[|Y^e - X^e b|^2] = \text{ causal parameter,} \tag{4}$$

where $\mathcal{F} = \{e;\ e \text{ satisfies the ad-hoc conditions 1 and 2}\}$. The definition of the causal parameter and the precise description of the result is given later in Section 3.1. The point here is to emphasize that the causal parameter or the causal solution is optimizing a certain worst case risk. This opens the door to think about causality in terms of optimizing a certain (worst case) risk. We believe that this is a very useful way which might ease some of the more complicated issues on structure search for causal graphs and structural equation models.

3 Invariance of conditional distributions

A key assumption for inferring causality from heterogeneous data as in (1) is an invariance assumption. It reads as follows:

(A($\mathcal{E}$)): There exists a subset $S^* \subseteq \{1, \ldots, p\}$ of the covariate indices (including the empty set) such that

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is the same for all } e \in \mathcal{E}.$$

That is, when conditioning on the covariates from S^* (denoted by $X^e_{S^*}$), the conditional distribution is invariant across all environments from $\mathcal{E}$.

$(\mathbf{A}(\mathcal{F}))$: Analogous but now for the much larger set of environments $\mathcal{F}$.

In a linear model setting, the invariance assumption translates as follows. There exists a subset S^* and a regression coefficient vector β^* with $\text{supp}(\beta^*) = \{j;\ \beta^*_j \neq 0\} = S^*$ such that:

$$\begin{aligned}\text{for all } e \in \mathcal{E}: \qquad & Y^e = X^e \beta^* + \varepsilon^e, \\ & \varepsilon^e \text{ independent of } X^e_{S^*},\ \varepsilon^e \sim F_\varepsilon,\end{aligned}$$

where F_ε denotes the same distribution for all ε^e. That is, when conditioning (regressing) on $X^e_{S^*}$, the resulting regression parameter and error term distribution are the same for all environments $e \in \mathcal{E}$. From a practical point of view, this is an interesting invariance or stability property and the set of covariates S^* plays a key component in "stabilizing across the environments", see also Figure 12 in Section 6.

If the invariance assumption holds, we are sometimes interested in describing the sets S^* which fulfill invariance. We then denote by

$(\mathbf{A}_S(\mathcal{E}))$: The subset S fulfills invariance saying that

$$\mathcal{L}(Y^e | X^e_S) \text{ is the same for all } e \in \mathcal{E}.$$

$(\mathbf{A}_S(\mathcal{F}))$: analogous but now for the set of environments $\mathcal{F}$.

When considering the invariance assumption for the set of unknown (future) environments $\mathcal{F}$, the sets S^* for which $(\text{A}_{S^*}(\mathcal{F}))$ holds are particularly interesting as they lead to invariance and stability for new, future environments which are not observed in the data. This is a key for solving worst case risk optimization with respect to a class of perturbations which can be arbitrarily strong as in (4).

3.1 Invariance and causality

A main question is whether there are sets S for which $(\text{A}_S(\mathcal{F}))$ holds and if so, whether there are many such sets and how one can describe them. Obviously, this depends on $\mathcal{F}$ and the problem then becomes as follows: under what model $\mathcal{F}$ can we have an interesting description of sets S which satisfy the invariance assumption $(\text{A}_S(\mathcal{F}))$.

To address this at least in part, we consider structural equation models (SEMs):

$$\begin{aligned}& Y \leftarrow f_Y(X_{\text{pa}(Y)}, \varepsilon_Y),\ \varepsilon_Y \text{ independent of } X_{\text{pa}(Y)}, \\ & X \sim F_X, \end{aligned} \tag{5}$$

where $\text{pa}(Y)$ denotes the parental set of the response variable Y in the corresponding causal influence diagram, and the distribution F_X of X can be arbitrary but assuming finite second moments and positive definite covariance matrix. Often in the literature, a SEM is considered for all the variables:

$$\begin{aligned}& Y \leftarrow f_Y(X_{\text{pa}(Y)}, \varepsilon_Y), \\ & X_j \leftarrow f_j(X_{\text{pa}(X_j)}, \varepsilon_j), \end{aligned} \tag{6}$$

with $\varepsilon_Y, \varepsilon_1, \ldots, \varepsilon_p$ mutually independent. The model in (6) is a special case of the SEM in (5), the former now assuming a structural equation part for the X-variables. Furthermore, the formulation in (5) also allows other hidden variables which may act on X but do not have a confounding effect on Y. The case of hidden confounders will be discussed later in Section 4.

The (direct) causal variables for Y are defined to be

$$S_{\text{causal}} = \text{pa}(Y).$$

The environments or perturbations e change the distributions of Y and X in model (5) and we denote the corresponding random variables by X^e and Y^e. The ad-hoc conditions 1 and 2 from Sections 2.3 are now formulated as follows:

(B($\mathcal{E}$)) The structural equation in (5) remains the same, that is for all $e \in \mathcal{E}$

$$Y^e \leftarrow f_Y(X^e_{\text{pa}(Y)}, \varepsilon^e_Y),\ \varepsilon^e_Y \text{ independent of } X^e_{\text{pa}(Y)},$$
$$\varepsilon^e_Y \text{ has the same distribution as } \varepsilon_Y.$$

(B($\mathcal{F}$)) analogous but now for the set of environments $\mathcal{F}$.

We note that the distributions of X^e are allowed to change.

The following simple result describes the special role of causality with respect to invariance.

Proposition 1. *Assume a partial structural equation model as in* (5). *Consider the set of environments $\mathcal{F}$ such that $(B(\mathcal{F}))$ holds. Then, the set of causal variables $S_{\text{causal}} = \text{pa}(Y)$ satisfies the invariance assumption with respect to $\mathcal{F}$, that is $(A_{S_{\text{causal}}}(\mathcal{F}))$ holds.*

The proof is trivial. The conditional distribution of Y^e given $X^e_{\text{pa}(Y)}$ is given by f_Y and the distribution F_ε of ε_Y, and these quantities do not depend on e. □

In presence of hidden confounder variables, invariance and causal structures can still be linked under certain assumptions: this will be discussed in Section 4. Proposition 1 says that causal variables lead to invariance: this has been known since a long time, dating back to Haavelmo (1943), see Figure 3. The result in Proposition 1 does not say anything about other sets of variables which satisfy the invariance assumption.

3.2 Invariant causal prediction

Roughly speaking, Haavelmo (1943) already realized that

$$\text{causal variables} \implies \text{Invariance}.$$

The reverse relation

$$\text{causal structures} \Longleftarrow \text{Invariance} \tag{7}$$

has not been considered until recently (Peters et al., 2016). This might be due to the fact that with nowadays large-scale data, it is much easier to infer invariance from data and thus, the implication from invariance to causal structures becomes much more interesting and useful.

The problem with the reverse implication (7) is the well-known identifiability issue in causal inference. We typically cannot identify the causal variables S_{causal}, unless we have very many environments (or perturbations) or making specific assumptions on nonlinearities (Hoyer et al.,

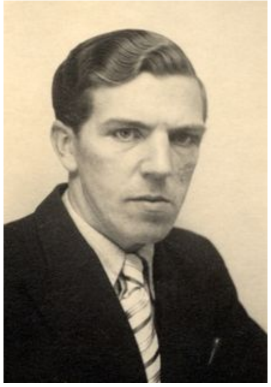


Figure 3: Trygve Haavelmo, Norwegian economist who received the Nobel Prize in Economic Sciences in 1989. Photo from `https://en.wikipedia.org/wiki/Trygve_Haavelmo`

2009; Bühlmann et al., 2014), non-Gaussian distributions (Shimizu et al., 2006) or error variances (Peters and Bühlmann, 2014). We will address the identifiability issue, which is often complicated in practice, in a fully "automatic" way as discussed next.

The starting point is to perform a statistical test whether a subset of covariates S satisfies the invariance assumption for the observed environments in $\mathcal{E}$. The null-hypothesis for testing is:

$$H_{0,S}(\mathcal{E}): \text{ assumption } (\mathrm{A}_S(\mathcal{E})) \text{ holds}$$

and the alternative is the logical complement, namely that assumption $(\mathrm{A}_S(\mathcal{E}))$ does not hold. It is worthwhile to point out that we only test with respect to the environments $\mathcal{E}$ which are observed in the data. To address the identifiability issue, we intersect all subsets of covariates S which lead to invariance, that is:

$$\hat{S}(\mathcal{E}) = \bigcap_S \{S;\ H_{0,S}(\mathcal{E}) \text{ not rejected by test at significance level } \alpha\}. \tag{8}$$

The specification of a particular test is discussed below in Section 3.2.2. The procedure in (8) is called Invariant Causal Prediction (ICP). The method is implemented in the `R`-package `InvariantCausalPrediction` for linear models (Meinshausen, 2018b) and `nonlinearICP` for non-linear models (Heinze-Deml and Peters, 2017), see also Section 3.2.2 below.

The computation of ICP in (8) can be expensive. There is an algorithm which provably computes ICP without necessarily going through all subsets (Peters et al., 2016): in the worst-case though, this cannot be avoided. If the dimension p is large, we advocate some preliminary variable screening procedure based on regression for the data pooled over all environments, see also Section 3.2.2 for the case with linear models below. Such regression-type variable screening procedures are valid when assuming a faithfulness condition: it ensures that the causal variables must be a subset of the relevant regression variables $\{j;\ X_j$ conditionally dependent of Y given $\{X_K; k \neq j\}\}$, see for example Spirtes et al. (2000).

We first highlight the property of controlling against false positive causal selections.

Theorem 1. *(Peters et al., 2016) Assume a structural equation model for the response Y as in* (5) *and that the environments or perturbations in $\mathcal{E}$ satisfy the assumption $(B(\mathcal{E}))$. Furthermore,* *assume that the tests used in* (8) *are valid, controlling the type I error. Then, for* $\alpha \in (0,1)$ *we have that*

$$\mathbb{P}[\hat{\mathcal{S}}(\mathcal{E}) \subseteq \mathrm{pa}(Y)] \geq 1 - \alpha.$$

The interesting fact is that one does not need to care about identifiability: it is addressed automatically in the sense that if a variable is in $\hat{\mathcal{S}}(\mathcal{E})$, it must be identifiable as causal variable for Y, at least with controllable probability $1 - \alpha$ (e.g. being equal to 0.95). For example, even if the environments in $\mathcal{E}$ correspond to ineffective heterogeneities (e.g. no actual perturbations), the statement is still valid.

Theorem 1 does not say anything about power. The power depends on the observed environments $\mathcal{E}$, besides sample size and the choice of a test. Roughly speaking, the power increases as $\mathcal{E}$ becomes larger: the more heterogeneities or perturbations, the better we can identify causal effects and this is also true for the procedure in (8). In fact, Peters et al. (2016, Th.2) discuss cases where the ICP method in (8) is able to identify the all the causal variables, i.e., where $\hat{\mathcal{S}}(\mathcal{E}) = \mathrm{pa}(Y)$ asymptotically as sample size tends to infinity. These special cases, where essentially all the variables are perturbed, are far from a complete understanding of necessary and sufficient conditions for identifiability of the causal variables. Furthermore, the construction with the intersection in (8) might be often conservative. In terms of power, one wants to reject as many sets of covariates which are violating invariance. This seems awkward at first sight but the fact that the tests are highly dependent helps to increase the probability that all sets S which do not fulfill the invariance hypothesis are rejected. A more quantitative statement of the latter and of power properties for ICP in general is difficult.

3.2.1 Some robustness properties.

The ICP procedure exhibits two robustness properties (and a third one is mentioned in Section 3.2.2 below).

Hidden confounding variables. Even in presence of hidden confounding variables (as in the scenario of Figure 5), we have the following: assuming a faithfulness condition (Spirtes et al., 2000, cf.),

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \mathrm{an}(Y)] \geq 1 - \alpha,$$

where $\mathrm{an}(Y)$ denotes the ancestor variables of Y. The details are given in Peters et al. (2016, Prop.5). In practice, this is interesting as we would still pick up some variables which indirectly have a total causal effect on Y.

Direct effects of environments on Y*.* The ad-hoc conditions 1 and 2 in Section 2.3 or the condition (B($\mathcal{E}$)) are violated if the environments directly affect Y. With a faithfulness condition (Spirtes et al., 2000, cf.), we would then always infer that no set S would fulfill the invariance assumption and therefore, as sample size gets sufficiently large, rejecting $H_{0,S}(\mathcal{E})$ for all S, we would obtain that $\hat{\mathcal{S}}(\mathcal{E}) = \emptyset$. Therefore, even under violation of the assumption that the environments or perturbations should not act directly on Y, the ICP procedure gives a conservative answer and claims no variable to be causal. In the literature, this scenario is also known under the name of so-called invalid instrumental variables (Guo et al., 2018, cf.).

3.2.2 Concrete tests.

We first assume that the structural equation for Y in (5) is linear with Gaussian error:

$$Y = \sum_{j \in \mathrm{pa}(Y)} \beta_j X_j + \varepsilon_Y,\ \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$$

and ε_Y is independent of $X_{\mathrm{pa}(Y)}$. The invariance hypotheses in $H_{0,S}(\mathcal{E})$ then becomes:

$H_{0,S}(\mathcal{E})^{\mathrm{lin-Gauss}}$**:** for all $e \in \mathcal{E}$ its holds that,

$$\begin{aligned} &Y^e = X_S^e \beta_S + \varepsilon_S^e,\ \varepsilon_S^e \text{ independent of } X_S^e \text{ (the same } \beta_S \text{ for all } e \in \mathcal{E}),\\ &\varepsilon_S^e \sim F_{\varepsilon_S} \text{ (the same for all } e \in \mathcal{E}). \end{aligned}$$

Thanks to the Gaussian assumption, exact tests for this null-hypothesis exist, for example with the Chow test (Chow, 1960). This is implemented in the `R`-package `InvariantCausalPrediction` (Meinshausen, 2018b).

The variable pre-screening methods, mentioned above after the introduction of the ICP estimator (8), become also much simpler in linear models. One can use e.g. the Lasso (Tibshirani, 1996) on the data pooled over all observed environments and then employ the ICP estimator for all subsets of $\hat{S}_{\mathrm{Lasso}} = \{j;\ \hat{\beta}_{\mathrm{Lasso},j} \neq 0\}$. To justify this, one needs to establish that, under $H_{0,S}^{\mathrm{lin-Gauss}}$, $\mathbb{P}[S \subseteq \hat{S}_{\mathrm{Lasso}}] \to 1$ asymptotically: sufficient conditions for this are given in e.g. Bühlmann and van de Geer (2011).

When the true underlying model for the response in (5) is sufficiently nonlinear but if one uses ICP in (8) with invariance tests based on a mis-specified Gaussian linear model, it typically happens that no set S satisfies the invariance assumption resulting in $\hat{S}(\mathcal{E}) = \emptyset$. One could use instead a testing methodology for nonlinear and non-Gaussian models to infer whether a subset of variables fulfills the invariance assumption. A corresponding proposal is given in Heinze-Deml et al. (2018) with the accompanying `R`-package `nonlinICP` (Heinze-Deml and Peters, 2017).

3.2.3 Application: single gene knock-out experiments.

We briefly summarize here the results from an application to predict single gene interventions in yeast (*Saccharomyces cerevisiae*); for details we refer to Meinshausen et al. (2016). The data consists of mRNA expression measurements of 6170 genes in yeast. We have 160 observational measurements from wild-type yeast and 1479 interventional data arising from single gene perturbation, where a single gene has been deleted from a strain (Kemmeren et al., 2014). The goal is to predict the expression level of a new unseen gene perturbation, i.e., the potential outcome of a new unseen perturbation.

More specifically, and using the terminology of the framework outlined before, we aim to infer some of the (direct) causal variables of a target gene. Denote the gene expression measurements by $G_1, \ldots, G_{6170}$. We consider as a response variable Y the expression of the jth gene and the corresponding covariates X the expressions of all other genes:

$$\begin{aligned} Y &= G_j,\\ X &= (G_1, \ldots, G_{j-1}, G_{j+1}, \ldots, G_p). \end{aligned}$$

The index $j \in \{1, \ldots, 6170\}$ and the covariate dimension is $p = 6169$. The aim is now to infer $\mathrm{pa}(Y)$, assuming a linear structural equation model as in (6) but now with functions f_Y and f_X being linear.

We construct the environments in a crude way: $\mathcal{E} = \{1, 2\}$ where the labels "1" and "2" denote the 160 observational and the 1479 interventional sample points, respectively. Thus, we pool all the interventional samples into one environment since we have no replicates of single gene perturbations. Other pooling schemes could be used as well: it is typically only an issue of power how to create good environments while the type I error control against false positive causal selections (Theorem 1) is still guaranteed, see also Section 3.2.4 below.

For validation, we do a training-test data splitting with a K-fold validation scheme of the interventional data: that is, we use all observational and $(K-1)/K$ of the interventional data with $K = 3$ or 5. We only consider true strong intervention effects (SIEs) where an expression X_k has a strong effect on Y (the perturbed value X_k are outside of the observed data range).

The predictions are based on ICP (with Bonferroni correction due to using the ICP procedure many times, once for each gene being the response variable). One then finds that 8 significant genes at corrected significance level $\alpha = 0.05$ and 6 of them are true positive strong intervention effects (Peters et al., 2016). What sticks out is that only very few causal genes have been found: in a graph with 6170 nodes (corresponding to all the genes), only 8 significant directed edges are found. When prioritizing the most promising causal genes, Figure 4 describes ROC-type curves: here ICP is supplemented with stability selection (Meinshausen and Bühlmann, 2010) on top of it for creating a "stabilized" ranking.

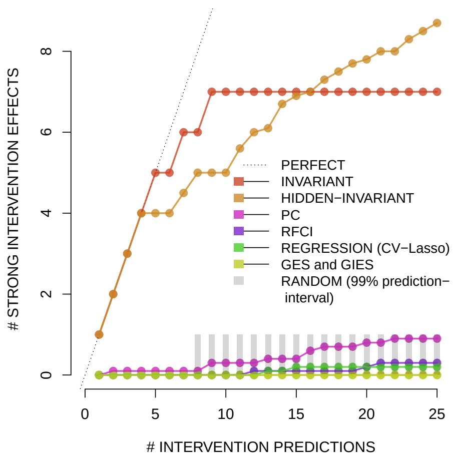


Figure 4: Prediction of strong intervention effects in single gene deletion experiments in Saccharomyces cerevisiae (yeast). x-axis: number of predictions made by a method; y-axis: number of predictions being true strong intervention effects.. Invariant causal prediction (ICP) in red: the first 5 predictions are all true (and 7 among the first 9 predictions are true). Orange: the Causal Dantzig Selector (Rothenhäusler et al., 2017), an algorithm based on an invariance property which includes hidden variables. All other methods are not distinguishable from random guessing (gray bars). Figure is taken from Meinshausen et al. (2016).

.

3.2.4 Unknown environments.

If the environments $e \in \mathcal{E}$ are not known, one can try to estimate them from data. The type I error control against false positive causal selections holds as long as the estimated partition $\hat{\mathcal{E}}$ does not involve descendant variables of the response Y: for example, one could use some clustering algorithm based on non-descendants of Y.

In practice, it is sometimes reasonable to assume that certain variables are non-descendants of Y. A canonical case is with time-sequential data: then, the environments can be estimated as different blocks of data at consecutive time points: this is some kind of a change point problem but now aimed for most powerful discovery with ICP. The methodology with time-sequential data is developed and analyzed in Pfister et al. (2018) and implemented in the `R`-package `seqICP` (Pfister and Peters, 2017).

4 Anchor regression: relaxing conditions

The main concern with ICP in (8) and the underlying invariance principle is the violation of the assumption in (B($\mathcal{E}$)), and thus also of the ad-conditions 1 and 2 from Section 2.3. Such a violation can happen under various scenarios and we mention a few in the following.

It could happen that only approximate instead of exact invariance holds. This would imply that we should only search for approximate invariance, something which we will incorporate in the anchor regression methodology described below in Section 4.2. Another scenario, say in a linear model, is that invariance occurs in the null-hypothesis $H_{0,S}(\mathcal{E})^{\text{lin}-\text{Gauss}}$ for the parameter $\beta_S^e \equiv \beta_S$ but with residual distributions ε_S^e which change for varying e; or vice-versa with invariant residual distribution but different regression parameters for varying e. This could be addressed in ICP by testing only either the parameter- or residual-part, where invariance is assumed to hold for the causal variables.

Perhaps the most prominent violation is in terms of hidden confounding variables H. The influence diagram from Figure 2 can then be extended to the situation of an instrumental variables (IV) regression model (Bowden and Turkington, 1990; Angrist et al., 1996; Rubin and Imbens, 2015), illustrated in Figure 5. Now, it is more convenient to think of the environments (or instruments) as random variables and we also model all random variables in the system in terms of a structural equation model (unlike as in (5), where we have only one structural equation for Y): instead of

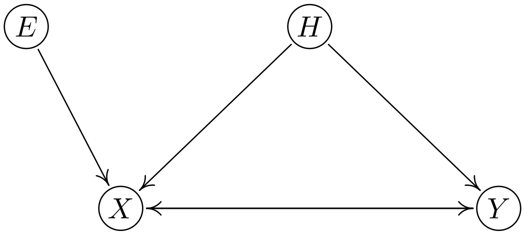


Figure 5: Graphical illustration with hidden confounding variables H. It corresponds to the instrumental variables regression model, where the instruments are now the environments.

(6), we consider now the IV regression model

$$Y \leftarrow f_Y(X_{\mathrm{pa}_X(Y)}, H, \varepsilon_Y),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}_X(X_j)}, H, E, \varepsilon_j),$$

where $H, E, \varepsilon_Y, \varepsilon_1, \ldots, \varepsilon_p$ are mutually independent of each other. The variable $H \in \mathbb{R}^r$ is hidden (not observed) and possibly confounding between X and Y (if some components of H are descendants of X or Y, these are not relevant for inferring the effect from X to Y and hence w.l.o.g. H is a source node). Here, $\mathrm{pa}_X(\bullet)$ denotes the parental variables X-variables (variables which are parents of $\bullet$ and from the set $\{X_1, \ldots, X_p\}$). The main assumption in the IV regression model requires that the instruments or environments do not directly influence the response variable Y nor the hidden confounders H (this is an extension of the ad-hoc conditions 1 and 2 from Section 2.3); and ideally, they would influence or change the X variables in a sufficiently strong way (as with the ad-hoc aim in Section 2.3). We are not going into more details from the vast literature on IV models with e.g. weak instruments, invalid instruments or partially identifiable parameters, see for example Stock et al. (2002); Murray (2006); Kang et al. (2016); Guo et al. (2018). Instead, we will relax this main assumption for an instrument as discussed next.

4.1 The anchor regression model

We will allow now that the environments can act directly also on H and Y, relaxing a main assumption in IV regression models. In the terminology of IV regression, we thus consider the case with so-called invalid instruments (Guo et al., 2018, cf.). This is an ill-posed situation for causal inference (from X to Y), yet it is still possible to obtain more meaningful results than what is obtained from standard regression methodology, see Section 4.3.2.

Instead of using the terminology "environment" we now us the word "anchor" (or anchor variable), for reasons which become more clear below. The structure of an anchor regression model is given by the graph in Figure 6. The anchor regression model, for simplicity here only in linear

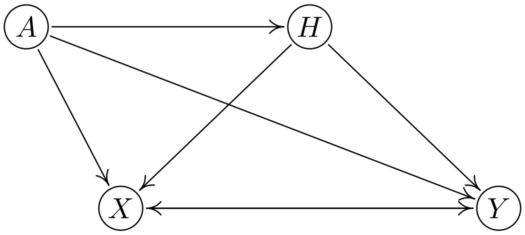


Figure 6: Graphical structure of the anchor regression model in (9), where A denotes the anchors which have been referred to as environments before. Note that the anchor variables are source nodes in the graph.

form, is defined as follows: it is a structural equation model of the form,

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA \tag{9}$$

We assume that all the variables are centered with mean zero. There could be feedback cycles and the graph of the structure could have cycles. In the latter case, we assume that $I - B$ is invertible (which always holds if the graph is acyclic). The main assumption is that A is a source node and thus, the contribution of A enters as an additional linear term MA. Because of this, we use the terminology "anchor": it is the anchor which is not influenced by other variables in the system and thus, it remains as the "static pole".

As mentioned above, one cannot identify the causal parameter $B_{Y,X}$, the row and columns corresponding to Y and X in the matrix B. However, we we will discuss in Section 4.3, that we can still get an interesting solution which optimizes a worst case risk over a class of scenarios or perturbations $\mathcal{F}$, using the terminology and spirit of (3).

4.2 Causal regularization and the anchor regression estimator

We are particularly interested in the structural equation for Y in the model (9) which we write as

$$Y = X^T\beta + H^T\alpha + A^T\xi + \varepsilon_Y,$$

with $X \in \mathbb{R}^p,\ H \in \mathbb{R}^q$ and $A \in \mathbb{R}^r$.

In the instrumental variables regression model where A would directly influence X only, it holds that H, ε_Y, A are mutually independent (but not so in the anchor regression model) and this then implies that

$$Y - X^T\beta \text{ is uncorrelated of } A.$$

Actually, we could substitute uncorrelatedness with independence in the IV model.

In cases where the causal parameter is non-identifiable, one could look for the solution

$$\text{argmin}_b \mathbb{E}[(Y - X^T b)^2] \text{ such that } \text{Corr}(A, Y - X^T b) = 0. \tag{10}$$

This leads to a unique parameter (assuming that $\text{Cov}(X)$ is positive definite), and there is a simple pragmatic principle behind it. This principle and the property of uncorrelatedness of the residual term with the anchor variables A also plays a key role in the anchor regression model for a class of so-called shift perturbations.

Similar to the idea in (10), we define the anchor regression estimator by using a regularization term, referred to as causal regularization, which encourages orthogonality or uncorrelatedness of the residuals with the anchor variables A. We denote the data quantities by the $n \times 1$ response vector $\mathbf{Y}$, the $n \times p$ covariate design matrix $\mathbf{X}$ and the $n \times q$ matrix $\mathbf{A}$ of the observed anchor variables. Let $\Pi_{\mathbf{A}}$ the projection in $\mathbb{R}^n$ onto the column space of $\mathbf{A}$. In practice, if the columns of $\mathbf{X}$ and $\mathbf{Y}$ are not centered, we would include an intercept column in $\mathbf{A}$. We then define

$$\hat{\beta}(\gamma) = \text{argmin}_b \left(\|(I - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2/n + \gamma\|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2/n \right). \tag{11}$$

We implicitly assume here that $\text{rank}(\mathbf{A}) = r < n$. For $\gamma = 1$, $\hat{\beta}(1)$ equals the ordinary least squares estimator, for $\gamma \to \infty$ we obtain the two-stage least squares procedure from IV regression and for $\gamma \to 0$ we adjust for the anchor variables in A. The properties of the anchor regression estimator in (11) are discussed next and make the role of the tuning parameter more clear.

The criterion function on the right-hand side of (11) is a convex function in b; for high-dimensional scenarios, we can add an ℓ_1-norm penalty, or any other sparsity inducing penalty:

$$\hat{\beta}(\gamma) = \text{argmin}_b \left(\|(I - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2/n + \gamma\|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2/n + \lambda\|b\|_1 \right). \tag{12}$$

The computation of the anchor regression estimator is trivial. We simply transform the variables Y and X,

$$\tilde{Y} = W_\gamma Y, \ \tilde{X} = W_\gamma X,$$
$$W_\gamma = I - (1 - \sqrt{\gamma})\Pi_{\mathbf{A}}.$$

The anchor regression estimator in (11) or (12) is then given by ordinary least squares or Lasso for the regression of $\tilde{Y}$ versus $\tilde{X}$.

4.3 Shift perturbations and robustness of the anchor regression estimator

The anchor regression estimator solves a worst case risk optimization problem over a class of shift perturbations.

We define the system under shift perturbations v by the same equations as in (9) but replacing the term MA from the contributions of the anchor variables by a deterministic or stochastic perturbation vector v. That is, the system under shift perturbations satisfies:

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} = B \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v = (I - B)^{-1}(\varepsilon + v).$$

The shift vector v is assumed to be in the span of M, that is $v = M\delta$ for some vector δ. The class of considered perturbations, denoted earlier as $\mathcal{F} \in (2)$, are shift perturbations as follows:

$$C_\gamma = \{v;\ \ v = M\delta \text{ for random or deterministic } \delta, \text{ uncorrelated with } \varepsilon$$
$$\text{and } \mathbb{E}[\delta\delta^T] \preceq \gamma\mathbb{E}[AA^T]\}. \tag{13}$$

Thus, C_γ contains shift perturbations whose length $\|v\|_2^2$ is typically $O(\gamma)$ as $\gamma \to \infty$.

For the case with $\gamma \to \infty$ one can characterize shift-invariance of residuals as follows.

Proposition 2. *(Rothenhäusler et al., 2018, Th.3) Assume that* $\mathbb{E}[AA^T]$ *is positive definite. Consider*

$$I = \{b \in \mathbb{R}^p;\ \mathbb{E}[A(Y - X^T b)] = 0\}.$$

Since Y and X have mean zero it follows also that $\mathbb{E}[Y - Xb] = 0$ and hence $\mathbb{E}[A(Y - X^T b)] = Corr(A, Y - Xb)$. Then,

$$b \in I \iff Y^v - (X^v)^T b \textit{ has the same distribution for all } v \in \text{span}(M).$$

With the goal to make the residuals invariant (for the class of shift perturbations), we aim to estimate the regression parameter β such that the residuals are encouraged to be fairly uncorrelated with A. This leads to the construction of the anchor regression estimator in (11) or (12).

A more general result than Proposition 2 is possible, uncovering a robustness property of the anchor regression estimator. We focus first on the population case. We denote by $P_A(\cdot)$ the projection operator, namely $P_A(Z) = \mathbb{E}[Z|A]$. In the anchor regression model (9) with $(I - B)$ being invertible, $P_A(Y)$ and $P_A(X)$ are linear functions in A. The population version of the anchor regression estimator is

$$\beta(\gamma) = \text{argmin}_b \left(\mathbb{E}[((I - P_A)(Y - X^T b))^2] + \gamma\mathbb{E}[(P_A(Y - X^T b))^2]\right). \tag{14}$$

Then, the following fundamental result holds.

Theorem 2. *(Rothenhäusler et al., 2018, Th.1) For any $b \in \mathbb{R}^p$ it holds that*

$$\sup_{v \in C_\gamma} \mathbb{E}[(Y^v - (X^v)^T b)^2] = \mathbb{E}[((I - P_A)(Y - X^T b))^2] + \gamma \mathbb{E}[(P_A(Y - X^T b))^2].$$

Thus, Theorem 2 establishes an exact duality between the causal regularized risk (which is the population version of the objective function for the estimator in (11)) and worst case risk over the class of shift perturbations. The regularization parameter equals the "strength" of the shift, as defined in (13): regarding its choice, see Section 4.3.1. A useful interpretation of the theorem is as follows. The worst case risk over shift perturbations can be considered as the one corresponding to future unseen data: this risk for future unseen data can be represented as a regularized risk for the data which we observe in the training sample. We further note that Theorem 2 holds for any b, and thus it also holds when taking the "argmin" on both sides of the equation. We then obtain that the population version $\beta(\gamma)$ in (14) is the minimizer of the worst case risk:

$$\beta(\gamma) = \text{argmin}_b \sup_{v \in C_\gamma} \mathbb{E}[(Y^v - (X^v)^T b)^2].$$

One can argue that also in the finite sample high-dimensional sparse scenario, the anchor regression estimator in (11) or (12) are asymptotically optimizing the worst case risk:

$$\sup_{v \in C_\gamma} \mathbb{E}[(Y^v - (X^v)^T \hat{\beta}(\gamma))^2] \leq \min_b \sup_{v \in C_\gamma} \mathbb{E}[(Y^v - (X^v)^T b)^2] + \Delta,$$

where, under suitable conditions, $\Delta \to 0$ as $n \to \infty$, or $p \geq n \to \infty$ in the high-dimensional scenario. The details are given in Rothenhäusler et al. (2018, Sec.4).

4.3.1 Choosing the amount of causal regularization.

The value of γ in the estimator (11) or (12) relates to the class of shift perturbations over which we achieve the best protection against the worst case, see Theorem 2. Thus, we could decide a-priori how much protection we wish to have or how much perturbation we expect to have in new test data.

Alternatively, we could do some sort of cross-validation. If the anchor variable encodes discrete environments, we could leave out data from one or several environments and predict on the left-out test-data optimizing the worst case error. If the anchor variables are continuous, the following characterization is useful.

The value γ in the causal regularization has also an interpretation as a quantile. Assuming a joint Gaussian distribution of the variables Y, X and A in the model (9), it holds that

$$\begin{aligned} & \alpha - \text{quantile of } \mathbb{E}[(Y - X^T b)^2 | A] \\ = \ & \mathbb{E}[((I - P_A)(Y - X^T \beta))^2] + \gamma \mathbb{E}[(P_A(Y - X^T \beta))^2], \\ & \text{for } \gamma = \alpha - \text{quantile of } \chi_1^2. \end{aligned} \tag{15}$$

The right-hand side also equals a worst case risk over shift perturbations, as stated in Theorem 2. Therefore, the relation above links the in-sample (for the non-perturbed data) quantiles to the out-sample (for the perturbed data) worst case risk. The exact correspondence of α and γ might not hold for more general situations. However, the qualitative correspondence is that a large α (high quantile) corresponds to a high value γ for the regularization term. Thus, one could choose a quantile value α, e.g. $\alpha = 0.95$, for the quantile of the conditional expectation of the squared error $\mathbb{E}[(Y - X^T b)^2 | A]$ and then calculate the γ which optimizes this quantile. Under a Gaussian assumption, we can estimate the quantities replacing expectations by mean squared test samples. This result also indicates that anchor regression with a large value of γ should result in good values for the high quantile of the squared prediction error (unconditional on A).

4.3.2 Diluted form of causality.

In the anchor regression model in general, it is impossible to infer the direct causal parameter β from X to Y. If the assumptions for instrumental variables regression are fulfilled, i.e., no direct effects from A to H and to Y and $\text{rank}(A) \geq \dim(X) = p$, then the anchor regression estimator with $\gamma \to \infty$ equals the unique two stage least squares estimator and consistently infers β; in particular, we also have that $\beta(\gamma \to \infty) = \beta$.

If the IV assumptions do not hold, for example in presence of invalid instruments where the anchor variables directly affect Y or H, the parameter $\beta(\gamma)$ with $\gamma \to \infty$ or γ being large is still a much more meaningful quantity than the standard regression parameter (with $\gamma = 1$). For large values of γ, the corresponding $\beta(\gamma)$ is minimizing a worst case risk over a class of large shift perturbations. This parameter and its entries with large values corresponding to important variables is interesting in many applications: the variables (with corresponding large parameter components of $\hat{\beta}(\gamma)$) are "key drivers" in a system of interest to explain the response Y in a stable manner over many perturbations. In fact, for $\gamma \to \infty$, we define

$$\text{supp}(\beta(\gamma \to \infty))$$

to be the set of variables which are "diluted causal" for the response Y (the variables which are relevant for Y in the framework of "diluted causality").

4.4 Some empirical illustrations

We illustrate the performance and behavior of the estimator (11) in the linear anchor regression model (9). We consider the case where the anchors are invalid instruments and low-dimensional and hence, inferring the causal effects from X to Y is impossible. The model for the variables A, H and X is as in model (M3) described later in Section 5.4, with $\dim(A) = r = 2$, $\dim(H) = q = 1$ and $\dim(X) = p = 10$. The structural equation for the response is

$$Y \leftarrow 3X_2 + 3X_3 + H - 2A_1 + \varepsilon_Y, \ \varepsilon_Y \sim \mathcal{N}(0, 0.25^2).$$

The training sample size is chosen as $n = 300$. The test sample is constructed with the same structure but now the anchor variables A are multiplied by the factor $\sqrt{10}$ and the test sample size is chosen as $n_{\text{out}} = 2000$. It is instructive to describe here how the anchor variables A act on X: the model is

$$X_j \leftarrow A_1\gamma_1 + A_2\gamma_2 + H + \varepsilon_{X_j}, \ \varepsilon_{X_j} \sim \mathcal{N}(0, 1),$$

where γ_1, γ_2 are coefficients which have been sampled i.i.d. from $\mathcal{N}(0, 1)$. The equation above changes in the test sample where we multiply A_1 and A_2 with the factor $\sqrt{10}$ which results in perturbations for the X_j variables. Figure 7 describes the quantiles of the absolute out-sample prediction error $|Y_{\text{out},i} - X_{\text{out},i}^T \hat{\beta}(\gamma)|$ where $\gamma = 7$ has been pre-specified. We also show the empirical

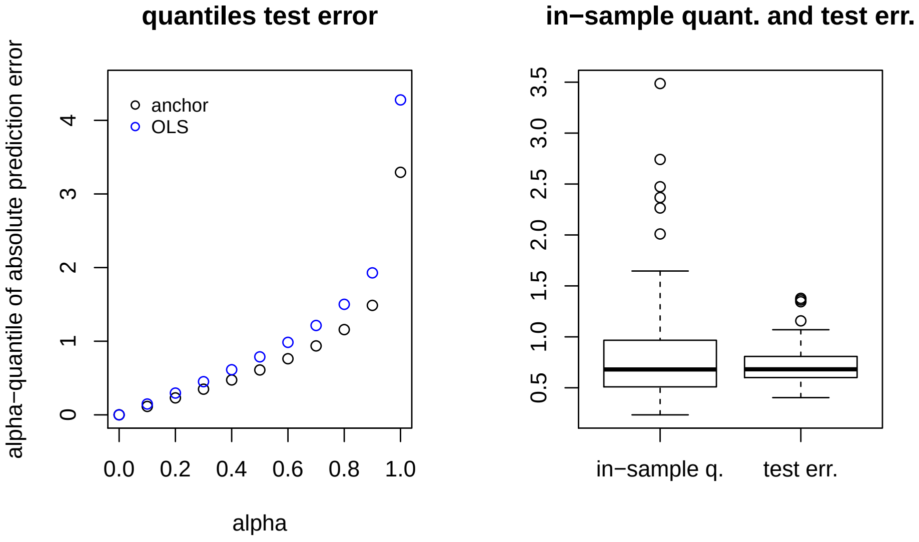


Figure 7: Left: empirical α-quantiles of $|Y_{\text{out},i} - \hat{Y}_{\text{out},i}|$ for $i = 1, \ldots, n_{\text{out}} = 2000$, averaged over 100 independent simulation runs. Right: estimated in-sample α-quantile of $\mathbb{E}[(Y - X^T\hat{\beta}(\gamma))^2|A]$ in (15) with $\alpha = 0.9918$ corresponding to $\gamma = 7$ (left boxplot) and out-sample (with perturbation) mean squared prediction error (right boxplot), for the 100 independent simulation runs.

relation between the in-sample quantile in (15) and the out-sample mean squared prediction error $\frac{1}{2000}\sum_{i=1}^{2000}(Y_{\text{out},i} - X_{\text{out},i}^T\hat{\beta}(\gamma))^2$. We conclude from the left panel of Figure 7 that the anchor regression estimator exhibits a substantially better prediction performance under perturbation out-sample scenarios than the ordinary least squares estimator. If the out-sample data would be generated as the in-sample training data, that is without new perturbations in the test data, there would be no gain, or actually a slight loss, of anchor regression over OLS (empirical results not shown here). Anchor regression only pays-off for prediction if some perturbations happen in new future data points which amplify the effect of heterogeneity (generated from the anchor variables A) in the future test data. This is briefly discussed next.

4.5 Distributional robustness

Anchor regression and causality can be viewed from the angle of distributional robustness (Heinze-Deml and Meinshausen, 2017; Meinshausen, 2018a). Distributional robustness refers to optimizing a worst case risk over a class of distributions:

$$\text{argmin}_\theta \max_{P \in \mathcal{P}} \mathbb{E}[\ell(Z;\theta)],$$

where $\ell(\cdot;\cdot)$ denotes a loss function, Z is the random variable generating a data point (e.g. $Z = (Y, X)$), θ is an unknown (potentially high- or infinite-dimensional) parameter and $\mathcal{P}$ is a class of probability distributions.

A typical choice for the class of distributions is

$$\mathcal{P} = \{P;\ d(P, P_0) \le \rho\},$$

where P_0 is the reference distribution, for example being the empirical measure, $d(\cdot,\cdot)$ is a metric, for example the Wasserstein distance, and ρ is a pre-specified radius, see for example Sinha et al. (2017); Gao et al. (2017).

For causality and anchor regression, the class of distributions $\mathcal{P}$ is given by a causal or "anchor-type" model consisting of perturbation distributions. Theorem 2 describes the connection more explicitly: the class $\mathcal{P}$ consists of amplifications of the observed heterogeneity in the data. This, because the perturbation distributions arise from shifts $v \in \mathrm{span}(M)$, thus being shifts in the direction of the effects from the observed anchor variable contribution which equals the term MA in the model (9); and the strength of the shifts in the perturbations is given by the parameter γ which has an analogous role as the radius ρ in the definition of $\mathcal{P}$ above. Thus, with anchor regression, the class of distributions is not pre-defined via a metric $d(\cdot,\cdot)$ and a radius ρ but rather through the observed heterogeneities in $\mathrm{span}(M)$ and a strength of perturbations or "radius" γ.

5 Nonlinear anchor regression

We present here a methodology for anchor regression generalizing the linear case. A core motivation is to design an algorithm for which any "machine learning" technology for regression can be plugged-in, including for example Random Forests or even Deep Neural Nets. We will argue that in presence of heterogeneity, essentially "any" of these machine learning methods can be improved using additional causal regularization.

We consider a nonlinear anchor regression structural equation model where the dependence of Y on X is a nonlinear function:

$$\begin{aligned} X &\leftarrow M_X A + B_{X,H} H + \varepsilon_X, \\ Y &\leftarrow f(X) + M_Y A + B_{Y,H} H + \varepsilon_Y, \\ H &\leftarrow M_H A + \varepsilon_H. \end{aligned} \qquad (16)$$

We can consider more general functions although a too high degree of nonlinearity can become more difficult with the anchor regression algorithm presented below. This is discussed after Corollary 1. We note that with M_Y and M_H being equal to zero we have a nonlinear instrumental variables regression model with linear dependences on the instruments (or the anchors) A and the hidden variables H.

5.1 The objective function and the algorithm

Consider a nonlinear regression function f, defined as

$$f(x) = \mathbb{E}[Y|X = x]$$

which is a map from $\mathcal{X}$ to $\mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the domains of X and Y, respectively. Given data $(Y^{(1)}, X^{(1)}), \ldots, (Y^{(n)}, X^{(n)})$, many estimation methods or algorithms for f can be written in the form

$$\hat{f} = \mathrm{argmin}_{f \in C} \frac{1}{2n} \sum_{i=1}^{n} (Y^{(i)} - f(X^{(i)}))^2 = \mathrm{argmin}_{f \in C} \frac{1}{2n} \|Y - f\|_2^2 \qquad (17)$$

where C denotes a suitable sub-class which incorporates certain restrictions such as smoothness or sparsity. On the right-hand side we have used a slight abuse of notation where $f = (f(X^{(1)}), \ldots, f(X^{(n)}))^T$ denotes the vector of function values at the observed $X^{(1)}, \ldots, X^{(n)}$. As will be seen below, the estimation algorithm is not necessarily of the form as in (17) but we use this formulation for the sake of simplicity.

Using the abbreviated notation with f mentioned above, the nonlinear anchor regression estimator is defined as:

$$\begin{aligned} &\hat{f}_{\text{anchor}} = \text{argmin}_{f \in C} G(f), \\ &G(f) = G_\gamma(f) = \frac{1}{2}\left(\|(I - \Pi_{\mathbf{A}})(Y - f)\|_2^2/n + \gamma\|\Pi_{\mathbf{A}}(Y - f)\|_2^2/n\right). \end{aligned} \tag{18}$$

As in (11), $\Pi_{\mathbf{A}}$ denotes the linear projection onto the column space of the observed anchor variable matrix $\mathbf{A}$. If the anchor variables A have a linear effect onto X, Y and the hidden confounders H, it is reasonable to consider the estimator with the linear projection operator $\Pi_{\mathbf{A}}$. We will give some justification for it in Section 5.5.

It is straightforward to see that the objective function can be represented as

$$G(f) = \|W(Y - f)\|_2^2/(2n), \quad W = W_\gamma = I - (1 - \sqrt{\gamma})\Pi_{\mathbf{A}}.$$

5.2 Anchor Boosting: a "regularized" approximation of the estimator

The question is how to compute or approximate the estimator in (18). We aim here for a solution where standard existing software can be used,

Our proposal is to use boosting. For this, we consider the negative gradient

$$-\frac{\partial}{\partial f} G(f) = W^2(Y - f)/n$$

and pursue iterative fitting of the negative gradient. The negative gradient fitting is done with a pre-specified base learner (or "weak learner"): it is a regression estimator $\hat{f}_{U,X}$ based on input data (U, X) with U denoting a response vector (e.g. $U = Y$ corresponds to the estimator applied to the original data). This is the standard recipe of gradient boosting (Breiman, 1999; Friedman, 2001; Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007) and the method is summarized in Algorithm 1. The choice of the stopping iteration m_{stop} is discussed in Section 5.2.1. The

Algorithm 1 Anchor Boosting algorithm.

```
1: Initialize with $f^{[0]} \equiv 0$. Set $m = 0$.
2: Increase $m$ by 1: $m \leftarrow m + 1$.
3: Compute the pseudo-response $\tilde{Y} = W^2(Y - f^{[m-1]})/n$ which equals the negative gradient vector
   evaluated at $f^{[m-1]}$.
4: Compute the regression function estimator $\hat{f}_{\tilde{Y},X}$ from the base learner and up-date

       $f^{[m]} = f^{[m-1]} + \nu \cdot \hat{f}_{\tilde{Y},X}$,

   where $0 < \nu < 1$ is a pre-specified parameter. The default value is $\nu = 0.1$.
5: Repeat steps 2-4 until reaching a stopping iteration $m_{\text{stop}}$.
```

stopping iteration is a regularization parameter: it is governing a bias-variance trade-off, on top of the regularization of causal regularization from anchor regression which is encoded in the matrix $W = W_\gamma$.

From another view point for regularization, we can think of the regression estimator $\hat{f}_{\tilde{Y},X}$ as an operator B (when evaluated at the observed X; it is a "hat" operator):

$$B : (\tilde{Y}, X) \mapsto \hat{f}_{\tilde{Y},X}(\cdot).$$

Then, it is straightforward to see that

$$f^{[m]}(X_1), \ldots, f^{[m]}(X_n) = \big(I - (I - W^2 B)^m\big) Y,$$

that is, the boosting operator at iteration m is equal to $\big(I - (I - W^2B)^m\big)$. If W^2B has a suitable norm being strictly < 1, then there is geometrical convergence to the identity which would fit the data Y perfectly. This indicates, that we should stop the boosting procedure to avoid overfitting. However, especially for large values of γ in $W = W_\gamma$, the norm of W^2B will be larger than one and geometrical contraction, i.e. overfitting, to the response vector Y will not happen.

5.2.1 Some criteria for choosing the stopping iteration.

In connection with a Random Forests (Breiman, 2001) learner for the estimator $\hat{f}_{\tilde{Y},X}$ in step 5 of Algorithm 1 we found that we can choose the stopping iteration m such as to minimize the objective function $\|W_\gamma(Y - f^{[m]})\|_2^2$ or even overshooting the minimum by say 10%. Formally, the two stopping rules are:

$$m_{\text{stop}} = \text{argmin}_m \|W_\gamma(Y - f^{[m]})\|_2^2, \tag{19}$$

$$\begin{aligned} m_{\text{stop}} &= \text{argmax}_m \|W_\gamma(Y - f^{[m]})\|_2^2 \\ &\quad \text{such that } \|W_\gamma(Y - f^{[m]})\|_2^2 \le 1.1 \min_m \|W_\gamma(Y - f^{[m]})\|_2^2. \end{aligned} \tag{20}$$

In general, we propose a rule based on the following observation. Consider the population version of $W = W_\gamma$ and denote it by

$$R_\gamma = \text{Id} - (1 - \sqrt{\gamma}) P_A,$$

where $P_A(\cdot)$ denotes the best linear projection onto A. That is,

$$P_A(Z) = (\alpha^0)^T A,\ \alpha^0 = \text{argmin}_\alpha \mathbb{E}[(Z - \alpha^T A)^2] = \text{Cov}(A)^{-1}\text{Cov}(A, Z).$$

Thus, R_γ is a function of A and random.

The population version of the optimization is

$$\text{argmin}_{f \in C_X} \mathbb{E}[(R_\gamma(Y - f(X)))^2] = \text{argmin}_{f \in C_X} \mathbb{E}[(R_\gamma Y - R_\gamma f(X))^2],$$

due to linearity of R_γ; here C_X denotes the class of measurable functions of X. We decompose this problem into two parts:

$$\begin{aligned} &\mathbb{E}[(R_\gamma Y - R_\gamma f(X))^2] = \mathbb{E}[(R_\gamma Y - g_{\text{opt}}(X, A))^2] + \mathbb{E}[(g_{\text{opt}}(X, A) - R_\gamma f(X))^2], \\ &g_{\text{opt}} = \mathbb{E}[R_\gamma Y | X, A]. \end{aligned}$$

This motivates a finite sample criterion guarding against overfitting. Whenever we estimate $f(\cdot)$ by $\hat{f}(\cdot)$ with boosting as in Algorithm 1, the residual sum of squares $\|W_\gamma(Y - \hat{f}(X))\|_2^2$ should be at least as large as the residual sum of squares of $\|W_\gamma Y - \hat{g}_{\text{opt}}(X, A)\|_2^2$ of any good and reasonably tuned machine learning estimator $\hat{g}_{\text{opt}}$. That is, we should choose the number of boosting iterations such that it minimizes the objective function under the given constraint:

$$\begin{aligned} m_{\text{stop}} &= \min_m \|W_\gamma(Y - f^{[m]}(X))\|_2^2 \\ &\text{such that } \|W_\gamma(Y - f^{[m]}(X))\|_2^2 \geq \|W_\gamma Y - \hat{g}_{\text{opt}}(X, A)\|_2^2. \end{aligned} \tag{21}$$

This guards against overfitting and avoids choosing a too large boosting iteration. We could also modify the rule to choose m as in (20) under the constraint that $\|W_\gamma(Y - f^{[m]}(X))\|_2^2 \geq \|W_\gamma Y - \hat{g}_{\text{opt}}(X, A)\|_2^2$.

5.2.2 Random Forests learner with a linear model component.

Besides using Random Forests as a base learner $\hat{f}_{U,X}$ in the Anchor-Boosting Algorithm 1, we consider also a modification which fits a linear model first and applies Random Forests on the resulting residuals. This modification is denoted by "LM+RF", standing for Linear Models+Random Forests. The LM+RF procedure is built on the idea that the linear model part is the "primary part" and the remaining nonlinearities are then estimated by Random Forests. This base leaner is able to cope well with estimating partial linear functions.

The "LM+RF" algorithms is defined as follows. Given a response variable Y and some covariates X:

1. Fit a linear model of Y versus X, by default including an intercept. The fitted (linear) regression function is denoted by $\hat{f}_1$.

2. Compute the residuals from step 1, denoted by R. Fit a Random Forests of R (being now the response variable) versus X: the fitted regression function is denoted by $\hat{f}_2$.

3. The final estimator is $\hat{f}_1 + \hat{f}_2$.

The LM+RF base learner is typically outperforming plain Random Forests if the underlying regression function is an additive combination of a linear and a nonlinear function.

5.2.3 Plug-in of any "machine learning" algorithm.

Obviously, any "machine learning" regression technique can be used as base learner in the Anchor Boosting Algorithm 1. In addition, also the stopping rule in (21) involving an estimator $\hat{g}_{\text{opt}}(X, A)$ can be used with any reasonable regression algorithm.

5.3 Some empirical results

We consider the following structural equation model for the in-sample data used for training with sample size $n = 300$.

$$\begin{aligned} &A \sim \mathcal{N}_2(0, \mathrm{I}), \\ &H \sim \mathcal{N}_1(0, 1), \\ &X_j = A_1 + A_2 + 2H + \varepsilon_{X,j} \ (j = 1 \ldots, p), \ \varepsilon_X \sim \mathcal{N}_{10}(0, 0.5^2 \cdot \mathrm{I}), \\ &Y = f(X_2, X_3) - 2A_1 + 3H + \varepsilon_Y, \ \varepsilon_Y \sim \mathcal{N}_1(0, 0.25^2), \end{aligned} \tag{22}$$

where $A, H, \varepsilon_X, \varepsilon_Y$ are jointly independent. The dimensions of the variables are $\dim(A) = 2$, $\dim(H) = 1$, $\dim(X) = 10$ and $\dim(Y) = 1$. For the function $f(\cdot)$, we consider the following two models:

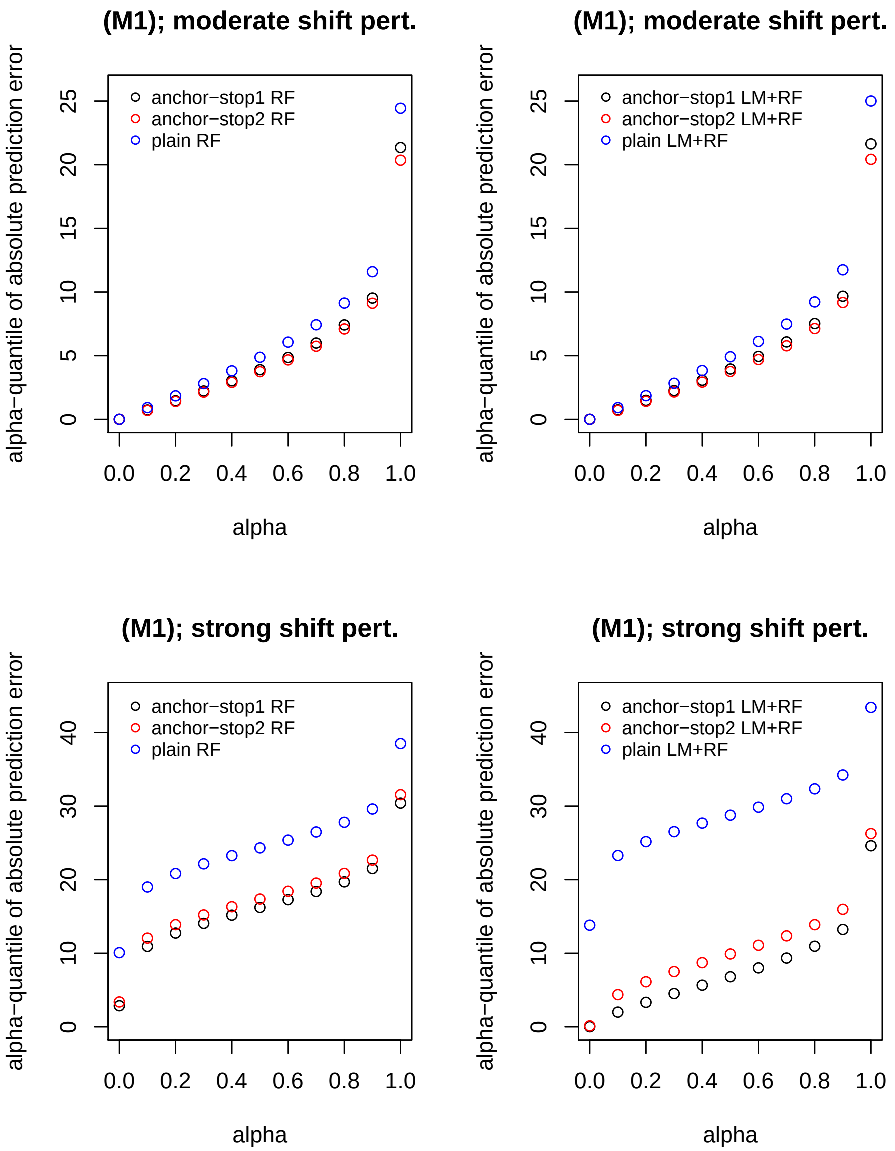


Figure 8: Empirical α-quantiles of $|Y_{\text{out},i} - \hat{Y}_{\text{out},i}|$ for $i = 1, \ldots, n_{\text{out}} = 2000$, averaged over 100 independent simulation runs. Model (M1) and moderate shift perturbations (i) for A_{out} (top) and strong shift perturbations (ii) for A_{out} (bottom). The Anchor Boosting algorithm is always used with $\gamma = 7$, with Random Forests (left) and with Linear Model + Random Forests (right), and with the two stopping criteria from (19) (stop1) and (20) (stop2).

(M1) $I(X_2 \leq 0) + I(X_2 \leq -0.5)I(X_3 \leq 1)$,

(M2) $X_2 + X_3 + I(X_2 \leq 0) + I(X_2 \leq -0.5)I(X_3 \leq 1)$.

The model (M1) has no linear term while (M2) does have one; for both models, the number of active variables is 2. The out-of-sample data is generated according to the same structural equation model as in (22) but with two different perturbations for the anchor variables, denoted by A_{out}. We consider the following:

(i) moderate shift perturbation:

$$\mu \sim \mathcal{N}_{\text{nout}}(1, 2^2\text{I}),$$
$$A_{\text{out}} \sim \mathcal{N}_{\text{nout}}(\mu, \text{I}),$$

where nout $= 2000$ denotes the number of out-of-sample observations.

(ii) strong shift perturbation:

$$\mu \sim \mathcal{N}_{\text{nout}}(10, \text{I}),$$
$$A_{\text{out}} \sim \mathcal{N}_{\text{nout}}(\mu, \text{I}),$$

where nout $= 2000$ denotes the number of out-of-sample observations.

We report some performances of Anchor Boosting with Random Forests, Anchor Boosting with the LM+RF learner from Section 5.2.2 and plain Random Forests in Figures 8-9. We do not tune the parameter γ and consider only the choice $\gamma = 7$. The performance measures are empirical α-quantiles of the out-of-sample predictions $|Y_{\text{out},i} - \hat{Y}_{\text{out},i}|$ for a range of different α-values. In terms of quantitative numbers, the relative performance gain of Anchor Boosting with stopping as in (20) ("stop2") over the corresponding plain Random Forests algorithm is given in Table 1.

model & learner	performance gains at $\alpha \in \{0.5, 0.8, 1\}$
(M1); moderate shift & RF	23.2%, 22.3%, 16.7%
(M1); moderate shift & LM+RF	23.5%, 22.7%, 18.3%
(M1); strong shift & RF	28.6%, 25.0%, 18.0%
(M1); strong shift & LM+RF	65.6%, 57.1%, 39.5%
(M2); moderate shift & RF	12.5%, 11.1%, -3.7%
(M2); moderate shift & LM+RF	19.5%, 18.9%, 12.6%
(M2); strong shift & RF	-100.8%, -73.9%, -30.1%
(M2); strong shift & LM+RF	83.3%, 73.4%, 47.6%

Table 1: Relative performance gain of empirical α-quantiles of absolute out-of-sample prediction errors (see also captions in Figures 8-9) of Anchor Boosting with $\gamma = 7$ and stopping from (20) over the corresponding plain base learner. The base learners are Random Forests (RF) and Linear Model + Random Forests (LM+RF).

The performance gain with Anchor Boosting is substantial, with the only exception of the case (M2) with strong shift perturbations (ii) and the Random Forest base learner (RF). This is a situation where a RF learner is "mis-specified" since it cannot capture well the linear function part

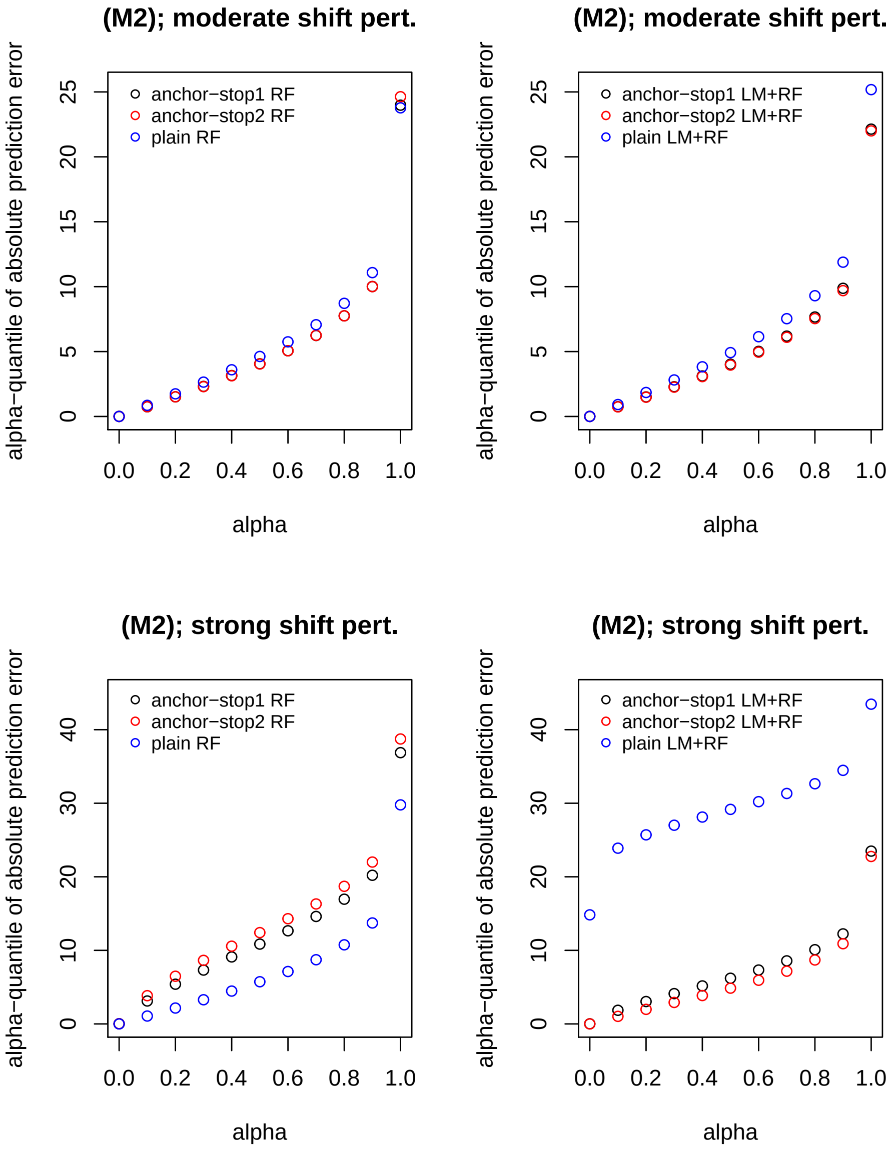


Figure 9: Empirical α-quantiles of $|Y_{\text{out},i} - \hat{Y}_{\text{out},i}|$ for $i = 1, \ldots, n_{\text{out}} = 2000$, averaged over 100 independent simulation runs. Model (M2) and moderate shift perturbations (i) for A_{out} (top) and strong shift perturbations (ii) for A_{out} (bottom). The Anchor Boosting algorithm is always used with $\gamma = 7$, with Random Forests (left) and with Linear Model + Random Forests (right), and with the two stopping criteria from (19) (stop1) and (20) (stop2).

in model (M2): it is particularly harmful with strong shift perturbations where a strong shift in the anchor variables A_{out} results in strong shifts of the covariates and through the linear function also in Y. The RF learner cannot capture such strong shifts well; when using the much better Linear

Model + Random Forests (LM+RF) learner, the performance gain of Anchor Boosting is massive. When comparing Anchor Boosting with LM+RF to plain Random Forests (RF), the relative gain over plain RF at $\alpha \in \{0.5, 0.8, 1\}$ in the model (M2) with strong strong shift perturbations is

$$15.2\%, 19.2\%, 23.6\%$$

which is again in clear favor of Anchor Boosting with the LM+RF learner.

Even if the model has no linear function (as in model (M1)), it seems to pay-off to use the LM+RF learner in presence of strong shift interventions. A possible reason is that extrapolation for large X-values is not easily possible with Random Forests.

5.4 Variable importance

Of particular interest is the notion of variable importance: in connection with (nonlinear) anchor regression, it is a more causally oriented measure than a plain variable importance measure in standard (nonlinear) regression. See also the term of "diluted causality" in Section 4.3.2.

We suggest a variable importance measure based on permutation, following Breiman's proposal for Random Forests (Breiman, 2001). Unlike Breiman's original proposal which involves the out-of-bag observations occurring in Random Forests, we work with the training data only. It seems to work as long as the estimated anchor regression function is reasonably regularized and does not overfit.

Consider an estimated anchor regression function $\hat{f}_{\mathrm{anchor}}(\cdot) : \mathcal{X} \to \mathcal{Y}$. For $j \in \{1, \ldots, p\}$, permute the jth covariate X_j (permute the sample indices) and denote this permuted variable by $X_{\mathrm{perm}(j)}$. We define the variable, for $j \in \{1, \ldots, p\}$

$$\tilde{X}^{(i)}_{\mathrm{perm}(j)} = (X_1^{(i)}, \ldots, X_{j-1}^{(i)}, X_{\mathrm{perm}(j)}^{(i)}, X_{j+1}^{(i)}, \ldots, X_p^{(i)})^T, \; i = 1, \ldots, n,$$

where the jth component is permuted relative to the observed variable $X^{(i)}$. We then compute the residual sum of squares

$$\mathrm{RSS}_j = n^{-1} \sum_{i=1}^{n} (Y^{(i)} - \hat{f}_{\mathrm{anchor}}(\tilde{X}^{(i)}_{\mathrm{perm}(j)}))^2.$$

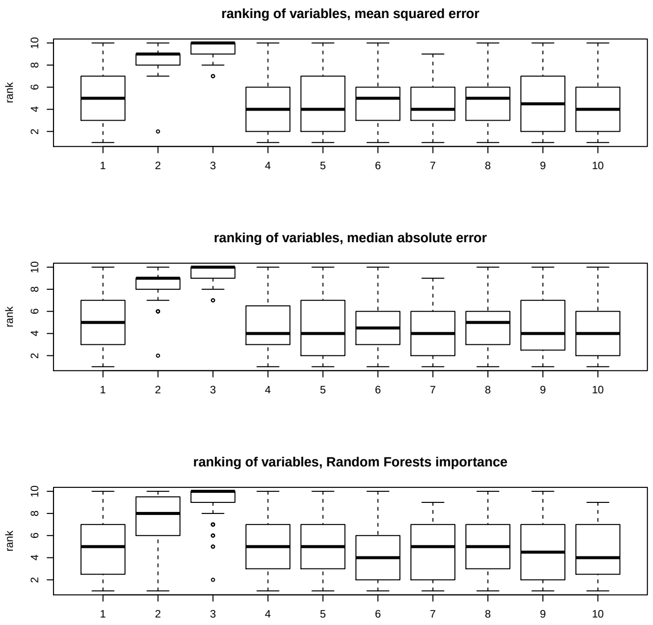
The importance measure is the relative increase in of RSS_j in comparison to the standard $\mathrm{RSS} = n^{-1} \sum_{i=1}^{n} (Y^{(i)} - \hat{f}_{\mathrm{anchor}}(X^{(i)}))^2$:

$$\mathrm{Imp}_j = \frac{\mathrm{RSS}_j - \mathrm{RSS}}{\mathrm{RSS}}. \tag{23}$$

As an alternative, we also consider the median absolute loss instead of the residual sum of squares:

$$\begin{aligned} \mathrm{Imp}_{\mathrm{med},j} &= \frac{\mathrm{h}_j - \mathrm{h}}{\mathrm{h}}, \\ h_j &= \text{sample median}\{|Y^{(i)} - \hat{f}(\tilde{X}^{(i)}_{\mathrm{perm}(j)})|;\; i = 1, \ldots, n\}, \\ h &= \text{sample median}\{|Y^{(i)} - \hat{f}(X^{(i)})|;\; i = 1, \ldots, n\}. \end{aligned} \tag{24}$$

For empirical results we note that the models (M1) and (M2) from Section 5.3 are very difficult in terms of identifying important variables. In fact, with the latter models, the correlation among


ranking of variables, mean squared error

ranking of variables, median absolute error

ranking of variables, Random Forests importance

Figure 10: Boxplots of the ranks of variable importance (rank 1 lowest priority for variable importance, rank 10 highest priority), based on 100 independent simulations of model (M3). The different boxes correspond to the different variable indices $j \in \{1, \ldots, 10\}$. Anchor Boosting with LM+RF and $\gamma = 7$: measure (23) (top), measure (24) (middle); variable importance of standard Random Forests measuring the increase in residual sum of squares based on out-of-bag observations (bottom).

the covariates X is extremely strong: the average absolute value of the off-diagonal elements of the empirical (in-sample) correlation matrix is found to be

$$\frac{1}{p(p-1)} \sum_{j \neq k} |\widehat{\mathrm{Corr}}(X_j, X_k)| = 0.97$$

for a representative sample. We modify to the following model:

(M3) The structural equation model is:

$$\begin{aligned}
&A \sim \mathcal{N}_2(0, \mathrm{I}),\\
&H \sim \mathcal{N}_1(0, 1),\\
&\Gamma \text{ a } 2 \times 10 \text{ matrix with i.i.d. } \mathcal{N}(0,1) \text{ entries},\\
&X_j = A^T \Gamma_{\bullet j} + H + \varepsilon_{X,j} \ (j = 1 \ldots, p),\ \varepsilon_X \sim \mathcal{N}_{10}(0, \mathrm{I}),\\
&Y = f(X_2, X_3) - 2A_1 + 3H + \varepsilon_Y,\ \varepsilon_Y \sim \mathcal{N}_1(0, 0.25^2),
\end{aligned}$$

where $A, H, \varepsilon_X, \varepsilon_Y, \gamma$ are jointly independent and

$$f(x_2, x_3) = x_2 + x_3 + I(x_2 \leq 0) + I(x_2 \leq -0.5) I(x_3 \leq 1).$$

The dimensions of the variables are $\dim(A) = 2$, $\dim(H) = 1$, $\dim(X) = 10$ and $\dim(Y) = 1$.

We consider again sample size $n = 300$ and out-of-sample size $n_{\text{out}} = 2000$ (for variable importance, we do not need this). In terms of the empirical performance for prediction, the analogue of Table 1 is as follows.

model & learner	performance gains at $\alpha \in \{0.5, 0.8, 1\}$
(M3); strong shift & RF	9.6%, 8.4%, 6.3%
(M3); strong shift & LM+RF	28.5%, 18.5%, 1.5%

For variable importance, the measures Imp_j and $\text{Imp}_{\text{med},j}$ are displayed in Figure 10. Since the variables X_2 and X_3 are the only active variables in the true function $f(X)$ in model (M3) we conclude that Anchor Boosting with LM+RF does a substantially better job for quantifying variable importance than standard Random Forests. Note that, as in Section 5.3, there was no tuning for the parameter γ which was set to the value $\gamma = 7$.

5.5 Some arguments why a simple linear projection $\Pi_{\mathbf{A}}$ is sufficient

We present here some arguments under what conditions a linear projection $\Pi_{\mathbf{A}}$ in the estimator in (18) is reasonable even in presence of a nonlinear function $f(\cdot)$.

We consider the situation when the regularization parameter $\gamma \to \infty$. Then, in the population version, the regularization enforces a solution $f(\cdot)$ with $\text{Corr}(A, Y - f(X)) = 0$. We thus consider the set

$$I = \{f(\cdot);\ \mathbb{E}[A(Y - f(X))] = 0 \text{ and } \mathbb{E}[Y - f(X)] = 0\}.$$

We then have the following result.

Proposition 3. *Consider a nonlinear anchor regression model as in* (16). *We assume that Cov(A) is positive definite. Assume that $\mathbb{E}[Y - f(X)|A] = \mu_f + \alpha_f^T A$ is a linear function of A with $\mu_f \in \mathbb{R}$ and α_f a $r \times 1$ vector. Then, for any $f \in I$ we have for every* do*-perturbation on A with* $\text{do}(A = a)$ *(Pearl, 2009, cf.), where the value a is deterministic or random:*

$$\mathbb{E}[Y^a - f(X^a)] = \mu_f = \textit{ constant w.r.t. } a.$$

Here, Y^a and X^a denote the random variables corresponding to the do-perturbation $\text{do}(A = a)$.

A proof is given at the end of the section. Proposition 3 leads to invariance of the first moment of the residuals $Y^a - f(X^a)$ but is not saying anything on higher moments or the invariance of the distribution as in Proposition 2.

For nonlinear $f(\cdot)$, the conditional expectation $\mathbb{E}[Y - f(X)|A]$ is typically non-linear in A, thus violating the assumption of Proposition 3. There is one very notable an relevant exception though: for discrete anchor variables $A \in \mathcal{A}$ and discrete space $\mathcal{A} = \{1, \ldots, m\}$ with labels $1, \ldots, m$, we can always write

$$\mathbb{E}[Y - f(X)|A = a] = \mu_f + \alpha_f^T A$$

where A is now with dummy-encoding with $r = (m - 1)$-dimensional representation Positive definiteness of $\text{Cov}(A)$ is ensured by assuming $0 < \mathbb{P}[A = k] < 1$ for all k. This leads to the following consequence:

Corollary 1. *Assume discrete anchor variables A with dummy encoding. We further assume that $0 < \mathbb{P}[A = k] < 1$ for all k. Then, for any $f \in I$ we have for every* do*-perturbation on A with* $\mathrm{do}(A = a)$*, where the value a is deterministic or random:*

$$\mathbb{E}[Y^a - f(X^a)] = \mu_f = \textit{ constant w.r.t. } a.$$

We note that a do-perturbation on A would simply change the value of the dummy-encoding: say $\mathrm{do}(A_j = 5)$ would imply a 10-fold effect if the observed $A_j = 0.5$.

We point out that we do not need a specific additive form in the structural equation model. However, Corollary 1 is only interesting if the set I is non-empty. This is ensured if the structural equation for Y is additive in X and H and the anchor A is an instrument only influencing X, i.e., for a broad class of nonlinear instrumental variable regression models which satisfies:

$$\begin{aligned} Y &\leftarrow f(X) + g(H, \varepsilon_Y), \\ X &\leftarrow h(A, H, \varepsilon_X), \end{aligned}$$

where $A, H, \varepsilon_X, \varepsilon_Y$ are mutually independent exogenous variables (source nodes in the graph), Obviously, $Y - f(X)$ is then independent of A and therefore $f \in I$. We do not elaborate more under what conditions I is non-empty: in the case of linear structural equations we note the result in Proposition 2.

For non-discrete anchor variables and where the conditional expectation $\mathbb{E}[Y - f(X)|A] = \mu_f + \alpha_f^T A + A^T \beta_f A + \ldots$, with β_f a $r \times r$ matrix, is nonlinear in A, the nonlinear anchor boosting algorithm leads to invariance of the linearized part:

$$\mathbb{E}[Y^a - f(X^a)] \approx \mu_f + \mathbb{E}[a^T \beta_f a] + \ldots$$

where the expression does not involve dependence on A through the linear part $\alpha_f^T a$. A linear approximation is more likely to be good if A has only a linear influence on X, H and Y and if $f(\cdot)$ is not too far away from a linear function. Thus, we would conclude that the nonlinear anchor boosting estimator leads to good predictive robustness if the direct effects of A are linear and the nonlinearity enters via a nonlinear function $f(\cdot)$ in the structural equation for Y. If the conditional expectation $\mathbb{E}[Y - f(X)|A]$ becomes highly non-linear, one would need a different penalization which could be of the form such as

$$\max_{g \in \mathcal{G}} \mathbb{E}[g(A)(Y - f(X))],$$

where $\mathcal{G}$ is a suitable class of functions.

Proof of Proposition 3. Since A is source node, the do-perturbation on A is the same as the conditional expectation:

$$\mathbb{E}[Y - f(X)|A = a] = \mathbb{E}[Y - f(X)|\mathrm{do}(A = a)] = \mathbb{E}[Y^a - f(X^a)],$$

where the last equality is just a definition.

We have for $f \in I$ that

$$\mathbb{E}[A(Y - f(X))] = 0.$$

Therefore by linearity of the conditional expectation we obtain

$$\begin{aligned} 0 &= \|\mathbb{E}[A(Y - f(X))]\|_2^2 = \|\mathbb{E}[A\mathbb{E}[Y - f(X)|A]]\|_2^2 = \|\mathbb{E}[A(\mu_f + \alpha_f^T A)\|_2^2 \\ &= \|\mathbb{E}[A\alpha_f^T (A - \mathbb{E}[A])]\|_2^2 = \|\mathbb{E}[(A - \mathbb{E}[A])\alpha_f^T (A - \mathbb{E}[A])]\|_2^2 \\ &= \|\mathbb{E}[(A - \mathbb{E}[A])(A^T - \mathbb{E}[A^T])]\alpha_f]\|_2^2 = \|\Gamma\alpha_f\|_2^2 = \alpha_f^T\Gamma^T\Gamma\alpha_f, \end{aligned}$$

where $\Gamma = \mathrm{Cov}(A)$ is positive definite. This implies that $\alpha_f \equiv 0$. Note that we have exploited above that $\mu_f = -\alpha_f^T\mathbb{E}[A]$ since $\mathbb{E}[(Y - f(X))] = 0$ for all $f(\cdot)$ in I. Therefore, we have that

$$\mathbb{E}[Y^a - f(X^a)] = \mathbb{E}[Y - f(X)|\mathrm{do}(A = a)] = \mathbb{E}[Y - f(X)|A = a] \equiv \mu_f,$$

and thus being constant w.r.t. a. □

5.5.1 Some empirical results.

To investigate the nature of discrete anchor variables, we consider a model similar to (M2) and sample sizes $n = 300$ and $n_{\text{out}} = 2000$ as before. The structural equation model is as for model (M2), except that the anchor variables and the structural equation for X are as follows.

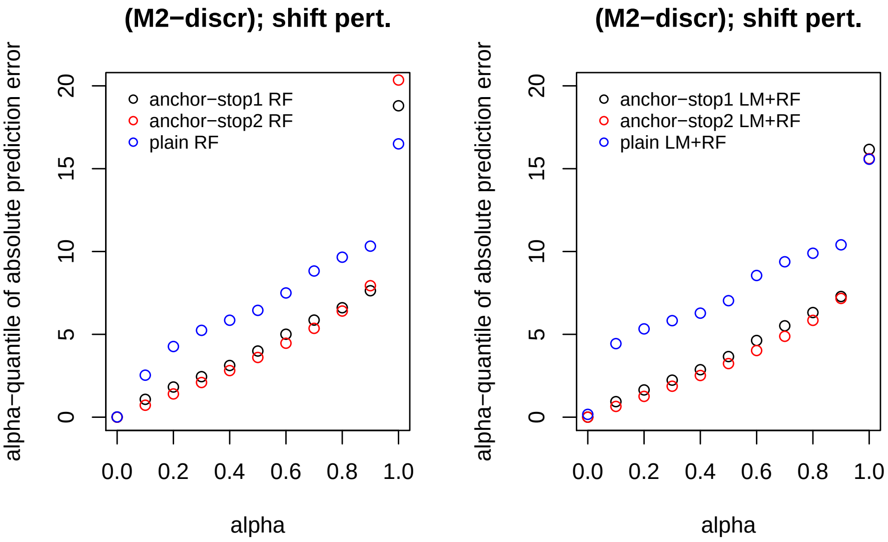


Figure 11: Empirical α-quantiles of $|Y_{\text{out},i} - \hat{Y}_{\text{out},i}|$ for $i = 1, \ldots, n_{\text{out}} = 2000$, averaged over 100 independent simulation runs. Model (M2-discr) with 3-fold amplification for A_{out}. The Anchor Boosting algorithm is always with $\gamma = 7$, with Random Forests (left) and with Linear Model + Random Forests (right), and with the two stopping criteria from (19) (stop1) and (20) (stop2).

(M2-discr)

$A_1 = (1, 1, \ldots, 1, 0, 0, \ldots, 0)^T$ where the first 150 entries are all 1 and then 0,
$A_2 = (0, 0, \ldots, 0, 1, 1, \ldots, 1)^T$ where the first 150 entries are all 0 and then 1.

The structural equation for X is

$$X_j^{(i)} = 2A_2^{(i)} - 2A_2^{(i)} + 2H^{(i)} + \varepsilon X, j^{(i)},\ \varepsilon_X \sim \mathcal{N}_{10}(0, 0.5^2 \mathrm{I}),$$

where $i = 1, \ldots, n$. The out-of-sample values of A_{out} are three times amplified:

$$A_1 = (3, 3, \ldots, 3, 0, 0, \ldots, 0)^T \text{ where the first 150 entries are all 3 and then 0,}$$
$$A_2 = (0, 0, \ldots, 0, 3, 3, \ldots, 3)^T \text{ where the first 150 entries are all 0 and then 3.}$$

The results are displayed in Figure 11 and are consistent with the earlier results in Figures 8-9.

6 Turning around the viewpoint

One can switch to an alternative view and, although perhaps thought-provoking, *define* a (diluted) form of causality via the invariance assumption.

Consider a class of perturbations $\mathcal{F}$. Invariance (of variables) with respect to the class of perturbations $\mathcal{F}$ is then defined as any set of covariates such that the invariance assumption in Section 3 holds, that is

$$\mathcal{L}(Y^e | X_S^e) \text{ is the same for all } e \in \mathcal{F}.$$

When $\mathcal{F}$ is sufficiently rich and fulfills the ad-hoc conditions 1 and 2 in Section 2 or the assumption (B($\mathcal{F}$)), then the $\mathcal{F}$ invariance corresponds to the causality in the literature: this is just another version of the worst case risk optimization viewpoint in (4). If $\mathcal{F}$ is not sufficiently rich, $\mathcal{F}$-invariance does not coincide with the set of causal variables, see also Figure 12.

If $\mathcal{F}$ violates the ad-hoc conditions, then invariance as above does not hold in general and is too demanding: but when restricting to shift perturbations v in an anchor regression model we still obtain invariance of the residuals

$$\mathcal{L}(Y^v - X^v \beta) \text{ is the same for all } v \in \mathcal{F},$$

where $\mathcal{F}$ is a subclass of shift perturbations, see Proposition 2 and formula (10). We referred to this as the $\mathcal{F}$ diluted causality (see Section 4.3.2) but we emphasize here that it is a certain shift invariance.

Such invariances have interesting implications in terms of interpretability and may lead to better insights in real applications. Thus, even for cases where inferring causality is ill-posed and non-identifiable, the invariance or diluted form of causality can provide more meaningful results and potentially contributes to an important aspect of "interpretable machine learning". We illustrate this point also in Figure 12.

7 Conclusions

Causality can be phrased in terms of worst case prediction risk, see Section 2.3, showing that causal inference is linked to a form of predictive robustness. The notion of (a certain) invariance can be beneficially exploited for predictive robustness and hence also for causality. The invariances themselves can be estimated from heterogeneous data: heterogeneity is important and

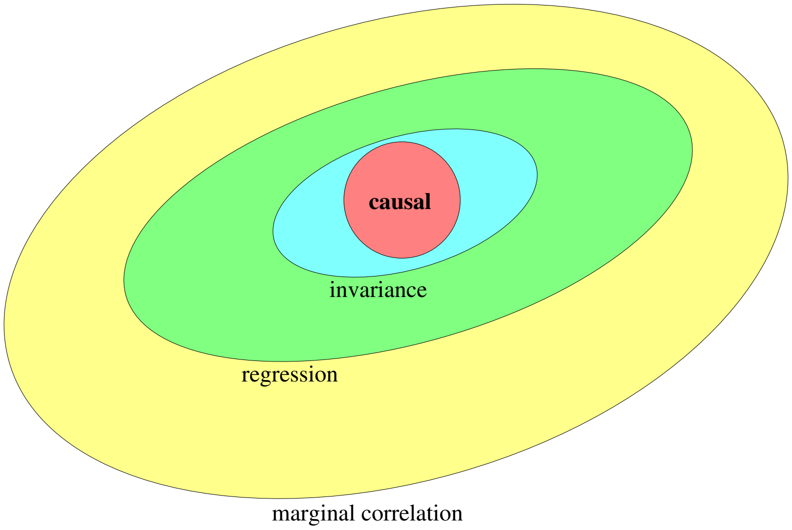


Figure 12: Illustration of various forms of associations between covariates X and a response Y. Marginal correlation of components of X with the response Y is a weak notion; collecting all variables with a non-zero regression coefficient is often more informative as it measures the partial correlation of components of X with Y. Under a faithfulness assumption, the causal variables are a subset of the relevant regression variables. In between is the notion of invariance and the diluted form of causality: even when inferring causality is impossible, obtaining the variables which satisfy invariance is often much more useful in many applications.

informative for inferring invariances. The paper includes a review of recent results, points to the R-packages `InvariantCausalPrediction` (Meinshausen, 2018b), `nonlinearICP` (Heinze-Deml and Peters, 2017) and `seqICP` (Pfister and Peters, 2017), and contributes some new developments for nonlinear problems in Section 5.

Our contribution can be seen as dealing with "statistics for perturbation (or heterogeneous) data". Even when causal inference is ill-posed, we show here some attempts to obtain predictive robustness and more meaningful approaches than what is provided by standard regression or curve fitting technology.

Acknowledgments

We thank Nicolai Meinshausen, Jonas Peters, Dominik Rothenhäusler and Niklas Pfister for many fruitful conversations.

References

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260.

Bowden, R. and Turkington, D. (1990). *Instrumental Variables*. Cambridge University Press.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science*, 22:477–505.

Bühlmann, P., Peters, J., and Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42:2526–2556.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, 98:324–339.

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

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95:407–424.

Editorial (2010). Cause and effect. *Nature Methods*, 7:243.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Gao, R., Chen, X., and Kleywegt, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48.

Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:793–815.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12.

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

Heinze-Deml, C. and Meinshausen, N. (2017). Conditional variance penalties and domain shift robustness. Preprint arXiv:1710.11469.

Heinze-Deml, C. and Peters, J. (2017). *nonlinearICP: Invariant Causal Prediction for Nonlinear Models*. R package version 0.1.2.1.

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60:578–586.

Hernán, M. A. and Robins, J. M. (2010). *Causal Inference*. CRC Boca Raton.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008)*, pages 689–696.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111:132–144.

Kemmeren, P., Sameith, K., van de Pasch, L., Benschop, J., Lenstra, T., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C., van Heesch, S., Kashani, M., Ampatziadis-Michailidis, G., Brok, M., Brabers, N., Miles, A., Bouwmeester, D., van Hooff, S., van Bakel, H., Sluiters, E., Bakker, L., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M., and Holstege, F. (2014). Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157:740–752.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248.

Maathuis, M., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37:3133–3164.

Meinshausen, N. (2018a). Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE.

Meinshausen, N. (2018b). *InvariantCausalPrediction: Invariant Causal Prediction*. R package version 0.7-2.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:417–473.

Meinshausen, N., Hauser, A., Mooij, J., Peters, J., Versteeg, P., and Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proc. National Academy of Sciences USA*, 113:7361–7368.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*, 20:111–132.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, second edition.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence interval (with discussion). *J. Royal Statistical Society, Series B*, 78:947–1012.

Pfister, N., Bühlmann, P., and Peters, J. (2018). Invariant causal prediction for sequential data. *Journal of the American Statistical Association, published online DOI 10.1080/01621459.2018.1491403*.

Pfister, N. and Peters, J. (2017). *seqICP: Sequential Invariant Causal Prediction*. R package version 1.1.

Pratt, L. Y. (1993). Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 204–211.

Richardson, T., Spirtes, P., et al. (2002). Ancestral graph markov models. *The Annals of Statistics*, 30:962–1030.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018). Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19:1309–1342.

Rothenhäusler, D., Bühlmann, P., and Meinshausen, N. (2017). Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics (to appear)*. arXiv preprint arXiv:1706.06159.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2018). Anchor regression: heterogeneous data meets causality. Preprint arXiv:1801.06229.

Rubin, D. and Imbens, G. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Shimizu, S., Hoyer, P., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Sinha, A., Namkoong, H., and Duchi, J. (2017). Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571.* Presented at Sixth International Conference on Learning Representations (ICLR 2018).

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search.* MIT Press, second edition.

Splawa-Neyman, J. ([1923] 1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translated and edited by D.M. Dabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczyc, Tom X (1923): 1–51 (Annals of Agricultural Sciences). *Statistical Science*, 5:465–472.

Stekhoven, D., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M., and Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28:2819–2823.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20:518–529.

Tchetgen, E. J. T. and VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21:55–75.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction.* Oxford University Press.