Anchor regression: heterogeneous data meet causality

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann and Jonas Peters

May 12, 2020

Abstract

We consider the problem of predicting a response variable from a set of covariates on a data set that differs in distribution from the training data. Causal parameters are optimal in terms of predictive accuracy if in the new distribution either many variables are affected by interventions or only some variables are affected, but the perturbations are strong. If the training and test distributions differ by a shift, causal parameters might be too conservative to perform well on the above task. This motivates anchor regression, a method that makes use of exogeneous variables to solve a relaxation of the causal minimax problem by considering a modification of the least-squares loss. The procedure naturally provides an interpolation between the solutions of ordinary least squares and two-stage least squares. We prove that the estimator satisfies predictive guarantees in terms of distributional robustness against shifts in a linear class; these guarantees are valid even if the instrumental variables assumptions are violated. If anchor regression and least squares provide the same answer (anchor stability), we establish that OLS parameters are invariant under certain distributional changes. Anchor regression is shown empirically to improve replicability and protect against distributional shifts.

1 Introduction

A substantial part of contemporaneous datasets are not collected under carefully designed experiments. Furthermore, data collected from different sources are often heterogeneous due to, e.g., changing circumstances, batch effects, unobserved confounders or time-shifts in the distribution. These heterogeneities or perturbations make it difficult to gain actionable knowledge that generalizes well to new data sets. Approaches to deal with inhomogeneities include robust methods [Huber, 1964, 1973], mixed effects models [Pinheiro and Bates, 2000], time-varying coefficient models [Hastie and Tibshirani, 1993, Fan and Zhang, 1999] and maximin effects [Meinshausen and Bühlmann, 2015].

On the other hand there is a growing literature on causal inference under various types of assumptions and different frameworks, with applications ranging from public health to biology and economics [Lauritzen and Spiegelhalter, 1988, Bollen, 1989, Greenland et al., 1999, Spirtes et al., 2000, Robins et al., 2000, Dawid, 2000, Rubin, 2005, Pearl, 2009, Peters et al., 2017]. Often the goal is to find the causes of some response variable Y among a given set of covariates X or to quantify the causal relationships between a set of variables. There are two main reasons why one is interested in the identification and quantification of causal effects. On one hand, it answers questions of the type "what happens to variable Y if we intervene on variable X", perhaps being the classical viewpoint of causality. On the other hand, predictions based on a causal model, that is, using the conditional mean of Y given all its causal predictors, will in general work equally well under arbitrary perturbations (interventions) on the covariates and thus, this provides an answer to the problem of generalization to new data sets mentioned above. The invariance property for prediction across interventions or perturbations has recently been exploited for causal inference [Peters et al., 2016] and a form of invariance plays a crucial role here as well.

In causal inference, one often considers so-called hard interventions that set some covariates to a certain value. In this paper, we instead consider interventions that shift the distribution of a target variable, which corresponds to an intervention on a variable that enters the target equation linearly. Using causal concepts for prediction under heterogeneous data seems attractive due to invariance guarantees under arbitrary shifts. In practice, however, exact invariance guarantees may be too conservative and can come with a price of subpar predictive performance on observational and moderately shifted data. We propose a balanced approach for trading off predictive performance on

observational data and predictive performance on perturbed (shifted) new data, with rigorous optimality guarantees under specific sets of perturbations or interventions. This can be cast as a form of distributional robustness, as discussed next. We consider being robust to interventional shifts in a particular class of models where identifying functionals yield a particularly elegant modification of OLS loss, with future work possibly allowing more general types of robustness to be developed. In addition to distributionally robust prediction, we will also consider the problem of distributionally robust variable selection. In this context, distributionally robust variable selection refers to the question whether a statistical parameter is invariant under certain distribution changes. Distributionally robust prediction and variable selection are closely related, as we will see below.

1.1 Distributionally robust prediction and variable selection

In a linear setting, the goal of distributionally robust prediction can be expressed as the optimization problem

$$\min_{b\in\mathbb{R}^d} \max_{F\in\mathcal{F}} \mathbb{E}_F[(Y - X^\intercal b)^2], \tag{1}$$

where X is a d-dimensional vector of covariates, Y is the target variable of interest, $\mathcal{F}$ is a class of distributions, and $\mathbb{E}_F$ takes the expectation w.r.t. $F \in \mathcal{F}$. Choosing different classes $\mathcal{F}$ results in estimators with different properties, see for example Sinha et al. [2018], Gao et al. [2017], Meinshausen [2018]. We first discuss two well-known choices of $\mathcal{F}$ and the corresponding estimators.

1.1.1 No perturbations and ordinary least squares

If $\mathcal{F}$ contains only the training (or observational) distribution, we write $\mathbb{E}_{\text{train}}$ and the optimization problem (1) becomes ordinary least squares,

$$b_{\text{OLS}} = \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2].$$

This does not take into account any distributional robustness. The sample version substitutes $\mathbb{E}_{\text{train}}$ by the sample mean over the observed data resulting in ordinary least squares estimation. We discuss in Section 1.3 that ℓ_2- and ℓ_1-norm regularized regression can also be derived from a sample version of (1) for a suitable class $\mathcal{F}$.

1.1.2 Intervention perturbations and causality

Assume now that the distribution (X, Y) is induced by an (unknown) linear causal model, e.g., a linear structural causal model, an example of which we will see in Section 2.1.If the class $\mathcal{F}$ contains all interventions on subsets of variables not including Y, then the optimizer of (1) is the vector of causal coefficients [e.g., Rojas-Carulla et al., 2018, Theorem 1]. That is,

$$b_{\text{causal}} = \operatorname*{argmin}_b \max_{F\in\mathcal{F}} \mathbb{E}_F[(Y - X^\intercal b)^2], \tag{2}$$

for $\mathcal{F}$ containing all interventions on (components) of X. Similarly, the causal parameters are optimal if in all distributions $F \in \mathcal{F}$ there are hard interventions on all parents and children of X (here, the interventions do not need to be arbitrarily strong). Both of these results are direct implications of well-known invariance properties of causal models [Haavelmo, 1944, Aldrich, 1989, Pearl, 2009].

In this spirit, a causal model can be seen as a prediction mechanism that works best under interventions on subsets of X that are arbitrarily strong or affect many variables. Under the training distribution, however, this solution is usually not as good as b_{OLS},

$$\mathbb{E}_{\text{train}}[(Y - X^\intercal b_{\text{causal}})^2] \geq \min_b \mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2] = \mathbb{E}_{\text{train}}[(Y - X^\intercal b_{\text{OLS}})^2], \tag{3}$$

with a potentially large difference. Hence in many cases, estimating the causal parameter leads to conservative predictive performance compared to standard prediction methods. The OLS solution on the other hand, can have arbitrarily high predictive error when the test distribution is obtained under an intervention.

This paper suggests a trade-off between these two estimation principles. Several relaxations of the problem in equation (2) are possible. Instead of protecting against arbitrarily strong interventions one can protect against interventions up to a certain size (norm). Also, perturbations in some directions may be more important than in other directions. Alternatively, instead of protecting against

interventions on all subsets of variables $X_1, \ldots, X_d$, one can attempt to find out which variables $S \subseteq \{X_1, \ldots, X_d\}$ are likely to be perturbed in the future. Then one can protect against interventions on the variables in S. For example, we might know (e.g., through background knowledge) that shifts in the distribution of X_1 are more likely than shifts in the distribution of X_2 on future data sets, which may be included in the class $\mathcal{F}$.

In this paper, we propose a new estimation principle, called *anchor regression*, see (4). We will see that under a linearity assumption, the proposed estimator can be written as a solution to (1), where the class $\mathcal{F}$ consists of certain shift interventions, i.e., interventions that shift numerical variables by a certain amount, which then propagate through the system.

1.1.3 Distributional replicability

Distributional replicability aims to understand whether a statistical parameter is stable under certain distributional changes. Replicability in this sense is distinctly different from statistical uncertainties due to finite samples, but closely related to the concepts of invariance and distributionally robust prediction. In the case of ordinary least-squares, it can be formalized as follows. The goal is to investigate whether

$$\operatorname*{argmin}_{b \in \mathbb{R}^d} \mathbb{E}_F[(Y - X^\intercal b)^2] \approx \operatorname*{argmin}_{b \in \mathbb{R}^d} \mathbb{E}_{F'}[(Y - X^\intercal b)^2],$$

for all $F, F' \in \mathcal{F}$, where $\mathcal{F}$ is some set of distributions. For example, two researchers may collect data about the same research question in different locations. Due to different circumstances, the data may come from two distributions $F \neq F'$. Even if the researchers use the same OLS model, they might get different estimates if the estimator is sensitive to small distributional changes.

We will see that *anchor regression* can be used to assess distributional replicability of OLS parameters across a certain set of distributions $\mathcal{F}$.

1.2 Our contribution

We propose an estimator that regularizes ordinary least squares with a penalty encouraging some form of invariance as mentioned above. The setting relies on the presence of exogenous variables which generate heterogeneity. We denote by $A \in \mathbb{R}^q$ such exogeneous variables and call them "anchors". If A is discrete, dummy encoding can be used in a pre-processing step to obtain $A \in \mathbb{R}^q$. Let X and Y be predictors and target variable, and assume that all variables are centered and have finite variance. Let further P_A denote the L_2-projection on the linear span from the components of A and write $\mathrm{Id}(Z) := Z$. We then define, for $\gamma > 0$, the solution b^γ to the population version of *anchor regression* as

$$b^\gamma := \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[((\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b))^2] + \gamma \mathbb{E}_{\text{train}}[(\mathrm{P}_A(Y - X^\intercal b))^2], \tag{4}$$

where $\mathbb{E}_{\text{train}}$ denotes the expectation over the observational or training distribution.

Turning to the finite-sample case, let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix containing observations of X. Analogously, the matrix containing observations of A is denoted by $\mathbf{A} \in \mathbb{R}^{n \times q}$, and the vector containing the observations of Y is denoted by $\mathbf{Y} \in \mathbb{R}^n$. We recommend a simple plug-in estimator for the *anchor regression* coefficient b^γ:

$$\hat{b}^\gamma = \operatorname*{argmin}_b \|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2, \tag{5}$$

where $\Pi_{\mathbf{A}} \in \mathbb{R}^{n \times n}$ is the matrix that projects on the column space of $\mathbf{A}$, i.e., if $\mathbf{A}^\intercal \mathbf{A}$ is invertible, then $\Pi_{\mathbf{A}} := \mathbf{A}(\mathbf{A}^\intercal \mathbf{A})^{-1} \mathbf{A}^\intercal$. Martin Emil Jakobsen realized that the family of finite sample estimators of anchor regression coincides with what is known as k-class estimators. These estimators have been suggested to improve IV-type estimation of structural parameters [Theil, 1958, Nagar, 1959]. In the high-dimensional case where $d > n$, an ℓ_1-penalty can be added to encourage sparsity. Computation of $\hat{b}^\gamma$ is simple as it can be obtained by running a least-squares regression of $\tilde{\mathbf{Y}} := (\mathrm{Id} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}})\mathbf{Y}$ on $\tilde{\mathbf{X}} := (\mathrm{Id} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}})\mathbf{X}$. More details on finite-sample *anchor regression* can be found in Section 4.

For $\gamma = 1$ we obtain the least squares solution, while for $\gamma > 1$ the *anchor regression* concept enforces that the projection of the residuals onto the linear space spanned by A is small ("near orthogonality"); the latter is related to the framework of instrumental variables regression. We will prove that the penalty term corresponds to the maximal change in expected loss under certain shift

interventions. In particular, we show that the solutions on the regularization path are optimizing a worst case risk under shift-interventions up to a given strength. In addition, we show that if *anchor regression* and ordinary least squares provide the same answer, the coefficients have a causal interpretation and are stable under certain distributional changes. More specifically, in this case the *anchor regression* coefficients are equal to OLS coefficients under certain perturbed distributions.

Under instrumental variables assumptions [Didelez et al., 2010], $b^{\infty} = b_{\text{causal}}$, i.e. one endpoint of anchor regression corresponds to the solution of equation (2). Our framework substantially relaxes the assumptions from the instrumental variables (IV) setting: in particular, we allow that the exogeneous anchor variables A are invalid instruments, as they are allowed now to directly influence (i.e., being direct causes of) Y or some hidden confounders H. The price to be paid for such cases is that the causal parameters are not identifiable any more. However, one can still exploit some invariance properties and obtain robust predictions in the sense of distributional robustness over a class $\mathcal{F}$ as introduced before. In addition, under the assumptions of instrumental variables, one can identify the causal parameters as the procedure naturally interpolates between the solution to ordinary least squares and two-stage least squares. One can also abandon causal and structural equation models and prove that the proposed anchor regression procedure minimizes quantiles of a conditional mean squared error.

The main benefits of the proposed *anchor regression* concept are robust predictions and replicability of variable selection on test data sets when the training data set can be grouped according to some exogeneous categorical variable (the "anchor") such as different circumstances, time-spans, experiments or experimental batches, or when certain numerical exogeneous variables are only available on the training, but not on the test data set. The anchor variable can either be used to encode heterogeneity "within" a data set or heterogeneity "between" data sets. More specifically, within one data set, each level of the anchor variable encodes a homogeneous group of observations of (X, Y). Alternatively, the anchor variable can be an indicator of data sets, where each data set is an homogeneous set of observations of (X, Y). In principle, it is possible to develop the theory for the case where the anchor is deterministic. However, for simplicity of exposition in this paper we will model the anchor variable as random.

Our *anchor regression* framework allows to quantitatively relate causality, invariance, robustness and replicability, under weaker assumptions than what is necessarily required to infer causal effects. Our work seems to be the first attempt to achieve this, with a practical procedure which is easy to compute and use in practice.

1.3 Related work

The considered perturbations from the class $\mathcal{F}$ are modeled by interventions in an underlying structural equation model [Pearl, 2009]. Furthermore, as the proposed procedure interpolates between the solution to ordinary least squares and the instrumental variables (two-stage least squares) approach, there are obvious connections to the IV literature, see e.g., [Wright, 1928, Bowden and Turkington, 1990, Didelez et al., 2010]. K-class estimators have the same algebraic form as anchor regression. The former are used to estimate structural parameters and often possess improved statistical properties compared to two-stage least squares, for example [Theil, 1958, Nagar, 1959]. In Leamer [1978] and Klepper and Leamer [1984] the authors show how backwards regressions can be used to bound the regression coefficients for errors-in-variables models. It is similar to our work in the sense that the considered model class forms a convex set, a structure which can be explored by modified linear regressions.

As mentioned above, predictive invariance in causal models has been exploited in Peters et al. [2016] for the purpose of learning direct causal effects. However in this work, the main goal is not to learn causal parameters, but to obtain predictive stability under perturbations. The goals of achieving robustness and learning causal parameters can be different, as shown by the example discussed in Section 2.2. In a different line of work, Pearl and Bareinboim [2014] have developed a formal language to treat the problem of generalizability of causal effects across environments or populations, assuming that the causal structure is known.

There exists a plethora of work on transfer learning in the machine learning literature, which focuses on knowledge transfer across different domains of the data [Pan and Yang, 2010]. Furthermore, there is work on distributional robustness, which explores bounded distributional perturbations, e.g., in a Wasserstein ball [Sinha et al., 2018] or under noise scaling [Heinze-Deml and Meinshausen, 2018]. In Rojas-Carulla et al. [2018] and Magliacane et al. [2018], the authors propose to use the best pre-

dictive model under all invariant models. In general, these methods do not allow for interventions on the target variable Y and concentrate on strong perturbations. Unlike pre-specifying the class $\mathcal{F}$, we aim to learn it from the training data: it has then the interpretation of an estimated class $\mathcal{F}$ which is generated from a structural equation model. Pfister et al. [2019] show for ODE based models that by trading off predictability and invariance under different experimental conditions in a similar way as *anchor regression*, one may still learn models that generalize better to unseen experiments. Yu and Kumbier [2020] expand traditional statistical uncertainty considerations by adding new notions of stability to improve reliability and reproducibility of knowledge extraction from data.

In Entner et al. [2013] the authors derive two rules that are sound and complete for inferring whether a given variable has a causal effect or not. The first rule uses (conditional) instruments to deduce the presence of a causal effect. While the goal of their work is different from the main intention of anchor regression, the first rule is similar to the condition that two version of anchor regression agree, as explained further below.

Furthermore, from a rather different viewpoint, it is known that many techniques for penalized regression can be formulated as a solution to (1), too. To see this, consider some measurement error ξ in X, i.e., that $(X+\xi, Y)$ under $\mathbb{P}_{\text{train}}$ has the same distribution as (X, Y) under $\mathbb{P}_{\text{test}}$. If we assume further that the measurement errors ξ_k are centered, jointly independent and independent of X and Y under $\mathbb{P}_{\text{train}}$, we can write

$$\mathbb{E}_{\text{test}}[(Y - X^\intercal b)^2] = \mathbb{E}_{\text{train}}[(Y - (X+\xi)^\intercal b)^2] = \mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2] + \sum_{k=1}^{d} \mathbb{E}_{\text{train}}[\xi_k^2] b_k^2.$$

If $\mathcal{F}$ contains all such test distributions with measurement errors up to strength $\mathbb{E}[\xi_k^2] \leq \gamma$, the optimization (1) becomes

$$\min_b \max_{F \in \mathcal{F}} \mathbb{E}_F[(Y - X^\intercal b)^2] = \min_b \mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2] + \gamma \sum_k b_k^2.$$

In words, under certain types of measurement errors, a (weighted) ridge penalty is optimal for prediction under perturbations. This is well known in the measurement errors literature, see for example Fuller [2009]. A similar result holds for the Lasso [Xu et al., 2009].

2 Population anchor regression

We now discuss properties of the population version of the proposed estimator (4). The overall goal is to predict the target variable $Y \in \mathbb{R}$ with the observed covariate vector $X \in \mathbb{R}^d$. The covariates X are potentially endogeneous, $A \in \mathbb{R}^q$ is a so-called anchor variable which is exogenous and $H \in \mathbb{R}^r$ is a vector of unobserved, or "hidden", random variables. In the case of categorical anchors, dummy encoding can be used to encode the categorical values with $A \in \mathbb{R}^q$.

To understand *anchor regression* and its properties, it is instructive to recognize the difference to the following well-known estimation concepts:

$$\begin{aligned}
b_{\text{PA}} &:= \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[((\text{Id} - \text{P}_A)(Y - X^\intercal b))^2] = \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[((Y - \text{P}_A Y) - (X - \text{P}_A X)^\intercal b)^2] \\
b_{\text{OLS}} &:= \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2] \\
b_{\text{IV}} &:= \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[(\text{P}_A(Y - X^\intercal b))^2] \\
b^\gamma &:= \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[((\text{Id} - \text{P}_A)(Y - X^\intercal b))^2] + \gamma \mathbb{E}_{\text{train}}[(\text{P}_A(Y - X^\intercal b))^2].
\end{aligned} \tag{6}$$

Here, PA stands for "partialling out", also sometimes called "adjusting for", and refers to linearly regressing out A from X and Y and considering residuals. The abbreviation IV refers to the two-stage-least-squares estimation principle in instrumental variable settings.

Due to the decomposition $\mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2] = \mathbb{E}_{\text{train}}[(P_A(Y - X^\intercal b))^2] + \mathbb{E}_{\text{train}}[((\text{Id} - P_A)(Y - X^\intercal b))^2]$, *anchor regression* coincides with ordinary least squares for $\gamma = 1$. For $\gamma = 0$, *anchor regression* coincides with b_{PA} and for $\gamma \to \infty$ it converges to b_{IV}, that is:

$$\begin{aligned}
b^0 &= b_{\text{PA}} \\
b^1 &= b_{\text{OLS}} \\
b^{\to\infty} := \lim_{\gamma \to \infty} b^\gamma &= b_{\text{IV}}.
\end{aligned} \tag{7}$$

The latter equation holds if b_{IV} is uniquely defined. Hence, *anchor regression* interpolates between b_{PA} and b_{OLS} for $0 \leq \gamma \leq 1$ and between b_{OLS} and b_{IV} for $1 \leq \gamma \leq \infty$.

Generally speaking, with *anchor regression*, we aim to learn a prediction mechanism that is reliable across A such as specific time periods, circumstances, locations or experimental batches observed in the training data set, and has some robustness guarantees regarding distributional shifts of observed and potentially also hidden variables. The structure of A crucially determines the robustness which we aim to achieve. For example, if we desire to achieve robustness across locations, then A should be chosen as a variable that encodes location in the training data set. If the desired robustness is with respect to experimental batches, then A should be chosen as a variable that describes different batches in the training data set.

While our estimator is defined under general conditions, most of our theoretical results focus on a model class that we introduce next.

2.1 A linear structural causal model

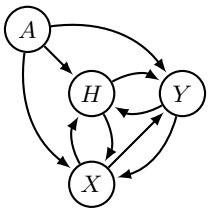
We assume that the data are generated from a linear structural equation model (SEM), also called a structural causal model (SCM), [Bollen, 1989, Pearl, 2009]. Let the distribution of (X, Y, H, A) under $\mathbb{P}_{\text{train}}$ be a solution of the SEM

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{M}A, \tag{8}$$

where $\mathbf{M} \in \mathbb{R}^{(d+1+r)\times q}$ and $\mathbf{B} \in \mathbb{R}^{(d+1+r)\times(d+1+r)}$ are unknown constant matrices and the anchors $A \in \mathbb{R}^q$, the hidden variables $H \in \mathbb{R}^r$, and the noise $\varepsilon \in \mathbb{R}^{d+1+r}$ are random vectors. We will call $\mathbf{M}$ the shift matrix. The random vectors A and ε are assumed to be independent. Furthermore, we assume that under $\mathbb{P}_{\text{train}}$, X and Y are centered to mean zero, that ε and A have finite second moments and that the components of ε are independent of each other. Equation (8) is potentially cyclic and a priori there may exist several or no distributions that satisfy this equation. In the following, we assume that $\text{Id} - \mathbf{B}$ is invertible. This guarantees that the distribution of (X, Y, H, A) under $\mathbb{P}_{\text{train}}$ is well-defined in terms of $\mathbf{B}$, ε, $\mathbf{M}$ and A as equation (8) has only one solution (equilibrium) satisfying

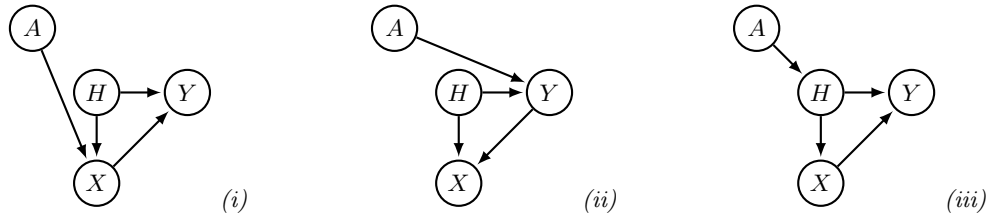
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (\text{Id} - \mathbf{B})^{-1}(\varepsilon + \mathbf{M}A).$$

More details on the interpretation in the cyclic case can be found in the Appendix, Section 8.1. The model induces a directed graph G, with the edges given by the following construction: For every $\mathbf{M}_{k,l} \neq 0$, a directed edge is drawn from A_l to the k-th variable in the $(d + 1 + r)$-dimensional vector (X, Y, H). Analogously, for every $\mathbf{B}_{k,l} \neq 0$, a directed edge is drawn from the l-th variable in (X, Y, H) to the k-th variable in (X, Y, H). The (vector-valued) variable A is called anchor since it corresponds to a source node in the directed graph, that is, there are no incoming edges into A. We allow the graph G to be cyclic. Note that the matrix $\text{Id} - \mathbf{B}$ is always invertible if the graph G is acyclic. An exemplary graph G that lies in our model class is given below. We also allow for self-cycles (for example an arrow from Y to Y), which are not depicted in the example.



Note that we do not assume A to be an instrument [Didelez et al., 2010]; we explicitly allow that A directly affects H and/or Y. This has important consequences: predictive guarantees of *anchor regression* do not exclusively apply to interventions on X but potentially also cover interventions on Y and H, depending on the data generating mechanism. More exemplary graphs and a potential motivation can be found in the following example.

Example 1 (Three examples of graphs G which are in our model class)**.** *Consider a setting with one-dimensional variables A, X and H. For example, X could be the activity of a certain gene, Y the activity of another gene and H the activity of a third, unobserved gene that regulates the activity of both X and Y. $A \in \{-1, 1\}$ could be an indicator variable of data collected from several experimental batches. The distribution of (X, Y, H) may change between the different batches $A \in \{-1, 1\}$. The change in distribution can be "caused" through a change in the activity of gene X (graph (i)), through a change in the activity of gene Y (graph (ii)) or a change in the activity of gene H (graph (iii)). Our model class contains many more graphs G than these three. Between the variables (X, Y, H) there are up to $3 \cdot 2 = 6$ directed arrows that may be in the graph (or not) and there are up to 3 arrows from A to (X, Y, H) that may be in the graph (or not), leading to a total of $2^3 \cdot 2^6 = 512$ directed graphs that lie in our model class for one-dimensional A, X and H.*



We aim to investigate the distribution of (X, Y, H) under perturbations. In the literature, so-called point, hard or do-interventions are often employed for causal modelling [Pearl, 2009].

Here, we aim to model the perturbed distributions as small, medium and potentially large perturbations of the training distribution. Interventions that act on the system in a linear fashion are often natural as well as simple to study. Thus, we will consider so-called shift interventions on (X, Y, H), which simply shift a variable by a value, see (9) below. This change subsequently propagates through the system. Shift interventions can be seen as a special case of a "parametric", "imperfect" or "dependent" intervention or a "mechanism change" [Eberhardt and Scheines, 2007, Korb et al., 2004, Tian and Pearl, 2001]. In particular, when A represents a "dummy encoding" of different batches, for example, we regard this as a flexible class of interventions.

The new interventional (perturbed) distribution is denoted by $\mathbb{P}_v$. The distribution of the variables (X, Y, H) under $\mathbb{P}_v$ is defined as the solution of

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + v, \tag{9}$$

where $v \in \mathbb{R}^{d+1+q}$ is a random or deterministic vector independent of ε, but not necessarily independent of A. The distribution of ε is assumed to be the same under $\mathbb{P}_{\text{train}}$ and under $\mathbb{P}_v$. We call v a shift. We potentially allow for interventions on X, Y and H, i.e., we allow $v_k \not\equiv 0$ for all $k \in \{0, \dots, d+q+1\}$. The main intuition behind shift interventions is that an external force shifts a certain variable by some amount. This shift propagates through the SEM, changing the distribution of some of the other variables.

2.2 Anchor regression: an example

First, we give an example of a linear SEM and the effect of a shift intervention. Then we will discuss the performance of ordinary least squares (OLS), the instrumental variables approach (IV) and partialling out A (PA); and motivate *anchor regression*. We compare the estimators by training them on the training distribution $\mathbb{P}_{\text{train}}$ and evaluating their performance on a perturbed distribution $\mathbb{P}_v$.

Consider a classical setting for the IV approach, where A is an instrument, X is endogenous and H is a hidden confounder. The structural equations of the unshifted distribution are defined on the left hand side of Example 2. The equations under a shift $v = (1.8, 0, 0)^\intercal$ are depicted on the right-hand side. The structural equations are assumed to be the same, but the variable X is shifted by $+1.8$ and the change propagates through the SEM.

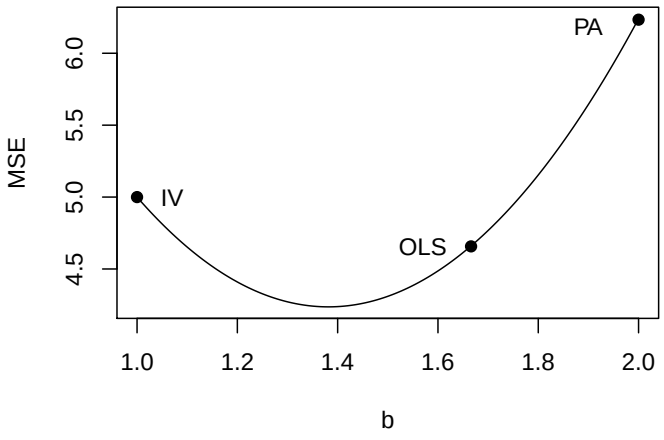


Figure 1: IV, OLS, PA and *anchor regression* coefficients are computed on unshifted data. The plot shows the MSE $\mathbb{E}_v[(Y - X^\intercal b)^2]$ on shifted variables for varying coefficients $b = b^\gamma$, $\gamma \in (0, \infty)$. The SEM for both shifted and unshifted data is given in Example 2. The optimal coefficient lies between IV and OLS.

Example 2. *The structural equations for $\mathbb{P}_{train}$ can be found on the left. On the right, structural equations for $\mathbb{P}_v$ with $v = (1.8, 0, 0)$.*

$$
\begin{aligned}
&A \sim \textit{Rademacher} \\
&\varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{indep.}{\sim} \mathcal{N}(0,1) \\
&H \leftarrow \varepsilon_H \\
&X \leftarrow A + H + \varepsilon_X \\
&Y \leftarrow X + 2H + \varepsilon_Y
\end{aligned}
\qquad\qquad
\begin{aligned}
&\\
&\varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{indep.}{\sim} \mathcal{N}(0,1) \\
&H \leftarrow \varepsilon_H \\
&X \leftarrow 1.8 + H + \varepsilon_X \\
&Y \leftarrow X + 2H + \varepsilon_Y
\end{aligned}
$$

There are two extreme cases for dealing with the variable A. The variation explained by A can be removed by partialling out A, sometimes also called residualizing with respect to A or adjusting for the effect from A. If we think about A as a subpopulation indicator variable, doing so creates a more homogeneous population and thus can correct for population stratification. The other extreme case is to remove all variation except for the variation explained by A. Under instrumental variables assumptions, doing so removes possible confounding variables and allows estimation of causal effects. For comparison, we thus consider partialling out the anchor variable (PA), ordinary least squares (OLS) and the instrumental variables approach (IV) in the form of two-stage least squares. All three are computed on $\mathbb{P}_{\text{train}}$, while their performance will be compared on the perturbed distribution $\mathbb{P}_v$.

If we regress Y on X, we obtain regression coefficient $b_{\text{OLS}} \approx 1.66$. The IV approach yields $b_{\text{IV}} = 1$ and partialling out A leads to $b_{\text{PA}} = 2$. For each coefficient $b^\gamma, \gamma \in [0, \infty)$ we compute the MSE on the shifted distribution $\mathbb{E}_v[(Y - X^\intercal b^\gamma)^2]$. The results are depicted in Figure 1. None of the three methods IV, PA and OLS yield the lowest MSE. In fact, large sections of the path of $b^\gamma, \gamma \in (1, \infty)$, outperform IV, PA and OLS. In that sense, even if IV regression identifies the true causal parameter, *anchor regression* can exhibit better prediction properties. This is not specific to the choice $v = (1.8, 0, 0)^\intercal$ but holds for other perturbations v as well. This will be discussed further in Section 2.4; it turns out that we can give optimality guarantees under certain interventions v, which depend on the underlying structural equation model. Furthermore, *anchor regression* will turn out to be useful even for cases where IV regression cannot identify the causal parameter, i.e., when the exogenous variable A is a direct cause of Y or the hidden confounder H. In the next section we discuss why all three approaches OLS, PA and IV have suboptimal performance in this example on the test data.

2.3 Trading off performance on perturbed and unperturbed data

Why did the three approaches OLS, IV and PA deliver suboptimal performance in the preceding example? Recall that the overall goal is to find b such that predictive performance is not only good on the training distribution but also under perturbed distributions. In this sense, we want to avoid "overfitting" to the particular distribution of the training data set. This can be investigated by considering the minimax loss

$$\operatorname*{argmin}_b \sup_{v \in C} \mathbb{E}_v[(Y - X^\intercal b)^2] \text{ for a suitable set } C \subseteq \mathbb{R}^{d+q+1}. \tag{10}$$

The crucial point here is to choose a "reasonable" set of perturbations C. If C is small, then the solution of equation (10) will usually not deliver good predictive performance under perturbations. If C is too large, then the solution may be unnecessarily conservative. Now let us return to the example of Section 2.2. It can be shown that b_{PA} solves the minimax problem for $C_{\text{PA}} = \{0\}$, i.e.,

$$b_{\text{PA}} = \operatorname*{argmin}_b \sup_{v \in C_{\text{PA}}} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

Hence it is not surprising that b_{PA} showed suboptimal performance under the intervention $v = (1.8, 0, 0)^\intercal$. Ordinary least squares solves the minimax problem for $C_{\text{OLS}} = \{v \in \mathbb{R}^3 : v_2 = v_3 = 0 \text{ and } v_1^2 \leq \mathbb{E}_{\text{train}}[A^2]\}$, i.e.,

$$b_{\text{OLS}} = \operatorname*{argmin}_b \sup_{v \in C_{\text{OLS}}} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

Loosely speaking, ordinary least squares optimizes the predictive performance under shifts in X up to strength $v_1^2 \leq \mathbb{E}_{\text{train}}[A^2]$. On the other hand, it can be shown that in the given example IV regression solves the minimax problem for $C_{\text{IV}} = \{v \in \mathbb{R}^3 : v_2 = v_3 = 0\}$:

$$b_{\text{IV}} = \operatorname*{argmin}_b \sup_{v \in C_{\text{IV}}} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

In words, the causal parameter (IV) solves the minimax problem if the supremum is taken over arbitrary strong shifts in X. Such shifts are not always realistic, hence from a prediction perspective the causal parameter can be unnecessarily conservative. The vector b_{PA} is optimized for prediction under zero perturbations $C_{\text{PA}} = \{0\}$ and does not exhibit stable predictive performance under shifts in X. As discussed earlier, ordinary least squares is somewhat in between.

The tradeoff is depicted in Figure 2: predictive performance of the four methods (PA, IV, OLS and *anchor regression* with $\gamma = 5$) is shown under varying intervention strength. While the causal parameter (IV) is the most stable, for small and medium-sized shifts other methods are preferable. On the other hand, OLS and PA show good performance only under small perturbations, with rapidly growing MSE for larger perturbations. Let $C^5 = \{v \in \mathbb{R}^3 : v_2 = v_3 = 0 \text{ and } v_1^2 \leq 5\}$. For the example it can be shown (cf. Theorem 1) that *anchor regression* for $\gamma = 5$ solves the minimax problem

$$\operatorname*{argmin}_b \sup_{v \in C^5} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

This gives us a convenient interpretation of b^γ for $\gamma = 5$: it minimizes the risk under shift interventions on X up to strength $|v_1| \leq \sqrt{5}$. The next section discusses the optimality of *anchor regression* under perturbations up to a given strength beyond the specific SEM of Example 2.

2.4 Optimal predictive performance under perturbations

In this section we will discuss a first main result, namely a fundamental connection between the population version of *anchor regression* and the worst case risk over a class of shift interventions. In Section 2.2 we saw that neither PA, OLS nor IV are optimal for prediction under the given intervention strength. The following theorem gives guarantees for the prediction error of *anchor regression* under shift interventions up to a given perturbation strength. Recall that P_A denote the L_2-projection on the linear span from the components of A. Under the assumptions of Section 2.1, we have $\mathrm{P}_A(X) = \mathbb{E}_{\text{train}}[X|A]$ and $\mathrm{P}_A(Y) = \mathbb{E}_{\text{train}}[Y|A]$. Let X and Y have mean zero.

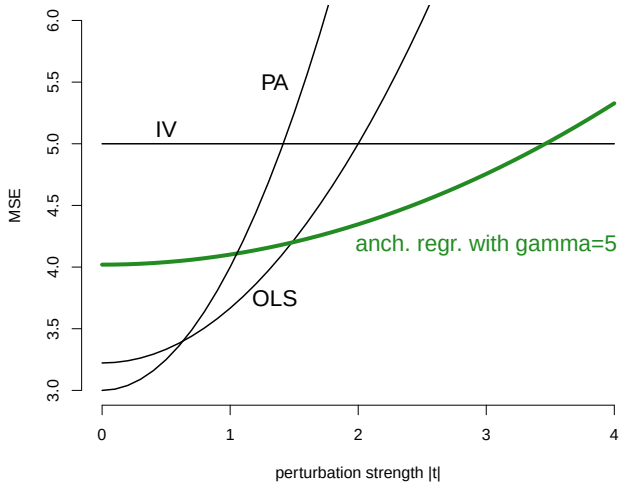


Figure 2: Predictive performance of the direct causal effect (IV), AP, OLS and *anchor regression* with $\gamma = 5$ under varying interventions on X. The SEM is taken from Example 2. The MSE $\mathbb{E}_v[(Y - X^\intercal b)^2]$ is depicted under perturbation strength $v = (t, 0, 0)^\intercal$. The causal parameter (IV) exhibits constant predictive performance under arbitrary perturbation strength $|t|$, but predictive performance under small perturbations is subpar. PA and OLS have very good performance under small interventions but performance suffers under larger interventions. *Anchor regression* with $\gamma = 5$ trades performance on unperturbed data ($t = 0$) for more stability, i.e., better performance on medium-sized interventions. In particular, it is minimax optimal under shifts $C^5 = \{(t, 0, 0)^\intercal : |t| \leq \sqrt{5} \approx 2.24\}$, cf. Theorem 1. For large shifts $|t|$ the IV method eventually outperforms *anchor regression*. Note that all shown solutions are anchor solutions, under respective penalties $\gamma = 0$ (PA), $\gamma = 1$ (OLS), $\gamma = 5$ and $\gamma = \infty$ (IV).

Theorem 1. *Let the assumptions of Section 2.1 hold. For any $b \in \mathbb{R}^d$ we have*

$$\mathbb{E}_{train}[((\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b))^2] + \gamma \mathbb{E}_{train}[(\mathrm{P}_A(Y - X^\intercal b))^2] = \sup_{v \in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2], \tag{11}$$

where

$$C^\gamma := \{v \in \mathbb{R}^{d+q+1} \text{ such that } vv^\intercal \preceq \gamma \mathbf{M} \mathbb{E}_{train}[AA^\intercal]\mathbf{M}^\intercal\}.$$

and $\mathbf{M}$ *is the shift matrix, cf. equation* (8)*. A formulation of the result where* v *is allowed to be random can be found in the Appendix, Section 8.5.*

Here, for two positive semidefinite matrices A and B we write $A \preceq B$ if and only if $B - A$ is positive semidefinite. In particular, we have $C^\gamma \subseteq \mathrm{span}(\mathbf{M})$. Readers familiar with the concept of interventions may thus think about $\mathbb{P}_v$ as the distribution under a point intervention on A, where the condition $v \in C^\gamma$ restricts the set of interventions to a certain strength.

There are two important takeaways from this theorem: First, the squared L_2-risk under certain worst-case shift interventions is equal to adding a penalty to the squared L_2-risk.

Second, as population *anchor regression* optimizes the penalized criterion (on the left-hand side of equation (11)), *anchor regression* minimizes the worst-case MSE under shift interventions up to a given strength in certain directions, cf. equation (6). We have discussed in Section 2.3 why it can be desirable to consider interventions only up to a given strength. In the following we want to briefly discuss the direction of the shift interventions in C^γ. To this end, note that

$$\mathrm{span}(\mathbf{M}) = \lim_{\gamma \to \infty} C^\gamma.$$

Here, for ease of interpretation we made the assumption that $\mathbb{E}_{\mathrm{train}}[AA^\intercal]$ is positive definite. We explicitly allow A to have a direct effect on X, Y or H. In other words, in the shift matrix $\mathbf{M}$, we allow $\mathbf{M}_{k\bullet} \not\equiv 0$ for some (or all) $k \in \{1, \dots, d + r + 1\}$. Hence C^γ potentially contains interventions that affect not only X but also Y or H. We discuss this in more detail in Section 8.2 in the Appendix.

Generally speaking, we have introduced a penalty that encourages good predictive performance under distributional shifts. Penalties of the form $\gamma\|b\|_2^2$ or $\gamma\|b\|_1$ are widely employed for finite sample regression to prevent overfitting the data with estimated parameters. Here, we deal with a different type of "overfitting" that may even affect the population version. For $\gamma = 0$ the population estimator will "overfit" to the particular distribution $\mathbb{P}_{\mathrm{train}}$, in the sense that it is not guaranteed to work well under shifted distributions $\mathbb{P}_v$. For $\gamma > 0$ we obtain predictive guarantees for both, shifted and unshifted data. As $\gamma \to \infty$, population *anchor regression* works increasingly well under strong interventions, at the price of deteriorating MSE on unshifted or moderately shifted data. In the finite sample case, additional regularization in form of an ℓ_1-penalty can be advisable. This is discussed in Section 4.2.

2.5 Limitations of using direct causal effects for prediction

In Section 2.2 we saw that using causal effects for prediction is in general not recommended if the perturbation strength is relatively small. In this section, we show that a similar caveat holds for the directions of the perturbations. Using direct (or total) causal effects in settings with perturbations on Y and H can be ill-advised, even if the perturbation strength is arbitrarily strong. Using direct causal effects for prediction does not protect against arbitrary perturbations.

As an example, consider the following structural equation model and a shift in the distribution of the hidden confounder H. On the left, the structural equation for the unperturbed distribution $\mathbb{P}_{\mathrm{train}}$ is defined. On the right, the data generating mechanism for the perturbed distribution $\mathbb{P}_v$ is given under a shift $v = (0, 0, t)^\intercal$, $t \in \mathbb{R}$.

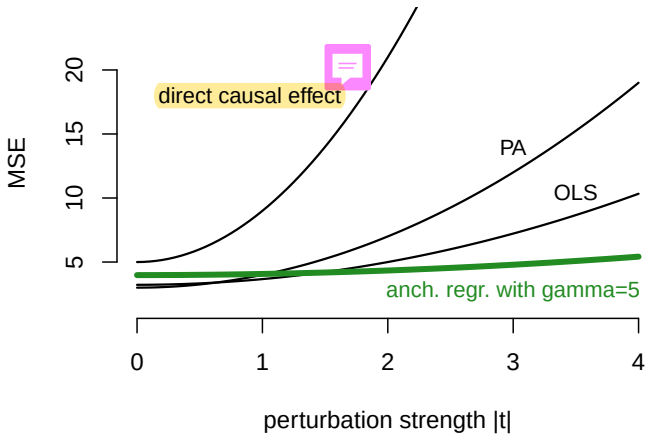


Figure 3: Predictive performance of the direct causal effect, PA, OLS and *anchor regression* under varying interventions on H. The MSE $\mathbb{E}_v[(Y - X^\intercal b)^2]$ is depicted under varying perturbations $v = (0, 0, t)^\intercal$. The corresponding structural equation models are given in equation (12). For small perturbations, PA and OLS perform better than *anchor regression*. The direct causal effect exhibits large MSE for all values of t. While the direct causal effect shows stable predictive performance under interventions on X (as discussed in Section 2.3), this is at the expense of predictive stability under interventions on H or Y. The MSE of *anchor regression* with $\gamma = 5$ slowly grows in $|t|$.

$$\begin{array}{c} A \sim \text{Rademacher} \\ \varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{indep.}}{\sim} \mathcal{N}(0,1) \\ H \leftarrow A + \varepsilon_H \\ X \leftarrow H + \varepsilon_X \\ Y \leftarrow 1 \cdot X + 2H + \varepsilon_Y \end{array} \qquad \begin{array}{c} \\ \varepsilon_H, \varepsilon_X, \varepsilon_Y \overset{\text{indep.}}{\sim} \mathcal{N}(0,1) \\ H \leftarrow t + \varepsilon_H \\ X \leftarrow H + \varepsilon_X \\ Y \leftarrow 1 \cdot X + 2H + \varepsilon_Y \end{array} \tag{12}$$

Assume that through some oracle (or previous experiments) we know that the direct causal effect from X to Y [Pearl, 2009, page 127] is 1, that is it equals the coefficient for X in the structural equation for Y. *Anchor regression* is trained on data from the SEM on the left; the predictive performance of *anchor regression* and the direct causal effect are compared on the shifted distribution $\mathbb{E}_v[(Y - X^\intercal b)^2]$. The results are shown in Figure 3. The direct causal effect is uniformly outperformed by PA, OLS and *anchor regression* with $\gamma = 5$. Roughly speaking, this is due to the fact that the direct causal effect is geared towards prediction under interventions on X, as discussed in Section 2.3. Interventions on H induce a very different distributional shift. Comparing PA and *anchor regression* leads to a similar conclusion as in Figure 2. Under small perturbations, PA and OLS are slightly better than *anchor regression*. However, *anchor regression* exhibits a stable performance across a large range of perturbation strengths and outperforms the other methods for medium or strong perturbations.

2.6 Interpretation of anchor regression via quantiles

We now provide an interpetation of *anchor regression* without using structural equation models. For reasons of simplicity, we present the result for continuous anchors. A similar result for discrete anchors can be found in the Appendix, Section 8.8. For the result of this section, the assumptions mentioned in Section 2.1 are not necessary, but instead we assume multivariate Gaussianity of (X, Y, A), see Lemma 1. Define $Q(\alpha)$ as the α-th quantile of $\mathbb{E}[(Y - X^\intercal b)^2|A]$. Recall that with the notation defined in Section 1.2 if (X, Y, A) is multivariate Gaussian we have $(\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b) = Y - X^\intercal b - \mathbb{E}[Y - X^\intercal b|A]$ and $\mathrm{P}_A(Y - X^\intercal b) = \mathbb{E}[Y - X^\intercal b|A]$.

Lemma 1. *Assume that the variables (X, Y, A) follow a centered multivariate normal distribution under $\mathbb{P}$. Then, for $0 \leq \alpha \leq 1$,*

$$Q(\alpha) = \mathbb{E}[((\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b))^2] + \gamma \mathbb{E}[(\mathrm{P}_A(Y - X^\intercal b))^2],$$

where γ equals the α-th quantile of a χ^2-distributed random variable with one degree of freedom.

Note that the right-hand side of the equation in Lemma 1 is the objective function of *anchor regression*. Thus, this shows that *anchor regression* can be used to optimize quantiles of $\mathbb{E}[(Y - X^\intercal b)^2|A]$, for example minimization of the 95%-quantile of $\mathbb{E}[(Y - X^\intercal b)^2|A]$ is achieved by b^γ with $\gamma = \chi^2_1(0.95)$. In spirit, this result is similar to Theorem 1. The perturbed distributions $\mathbb{P}_v$ in Theorem 1 play a similar role as the conditional distributions $\mathbb{P}[\bullet|A = a]$ in Lemma 1. For increasing γ, the predictions are increasingly reliable across distributions $\mathbb{P}[\bullet|A = a]$.

3 Replicability and Anchor Stability

We consider here the question of replicability when estimation is done a second time on a new perturbed dataset which has different data generating distributions than the original unperturbed but typically heterogeneous data. Replicability in this context is about potential differences in the regression parameters or prediction losses under different distributions: it is a "first order" problem instead of inferential statements about statistical uncertainties due to finite samples.

For the following two sections, we sometimes need a condition that the loss of anchor regression remains finite for $\gamma \to \infty$. We say the *projectability condition* is fulfilled if

$$\mathrm{rank}(\mathrm{Cov}_{\mathrm{train}}(A, X)) = \mathrm{rank}(\mathrm{Cov}_{\mathrm{train}}(A, X)|\mathrm{Cov}_{\mathrm{train}}(A, Y)), \tag{13}$$

where $\mathrm{Cov}_{\mathrm{train}}(A, X)|\mathrm{Cov}_{\mathrm{train}}(A, Y)$ is a $q \times (d + 1)$ matrix, consisting of the $q \times d$ covariance matrix $\mathrm{Cov}_{\mathrm{train}}(A, X)$, extended by the $q \times 1$ vector $\mathrm{Cov}_{\mathrm{train}}(A, Y)$. The reason why we call this the "projectability condition" becomes clear in Lemma 2 below.

The projectability condition (13) is fulfilled, for example, if $\mathrm{Cov}_{\mathrm{train}}(A, X)$ is of full rank and $q \leq d$ (sometimes called the under- or just-identified case as the dimension of A is less or equal to the dimension of X). The condition can also be fulfilled for $q > d$ under additional constraints on the nature of the link $A \to Y$. In general, the projectability condition allows that the anchor variables A directly influence also Y or H, and the example above for $q \leq d$ requires only a full rank condition on $\mathrm{Cov}_{\mathrm{train}}(A, X)$.

Lemma 2. *Assume that $\mathbb{E}_{train}[AA^\intercal]$ is invertible.*
The projectability condition (13) *is fulfilled if and only if*

$$\min_b \mathbb{E}_{train}[(\mathrm{P}_A(Y - X^\intercal b))^2] = 0. \tag{14}$$

The projectability assumption is testable in practice. The following results cover predictive stability and replicability under perturbations.

3.1 Replicability of the parameter $b^{\to\infty}$

Our first goal is to investigate the replicability of the parameter $b^{\to\infty}$. As stated in Theorem 1, this parameter vector is protecting against certain worst case shift perturbations of arbitrary strength and as such, it has an interesting interpretation; in analogy to causality which corresponds to worst case risk optimization for a different class of perturbations of arbitrary strength, see (2).

We consider two different data-generating distributions, and for notational coherence with before we denote them by "train" and "test". The training data is generated according to (8) and (9)

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + v, v = \mathbf{M}\delta, \ \delta = \kappa A + \xi, \tag{15}$$

where ξ is a random vector with mean zero and independent of ε and A and $\kappa \neq 0$. Note that with $\kappa = 1$ and $\xi = 0$ we have the model in (8).

The test data is from the following model:

$$\begin{pmatrix} X' \\ Y' \\ H' \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X' \\ Y' \\ H' \end{pmatrix} + \varepsilon' + v', v' = \mathbf{M}\delta',\ \delta' = \kappa' A' + \xi', \tag{16}$$

where ξ' is a random vector with mean zero and independent of ε' and A' and $\kappa' \neq 0$. We note that v' and A' can have arbitrarily different distributions than v and A but we assume that the dimensionalities are the same. The parameters **B** and **M** are the same in both models (15) and (16) and we assume that

$$\mathrm{Cov}_{\mathrm{test}}(\varepsilon') = L\mathrm{Cov}_{\mathrm{train}}(\varepsilon) \text{ for some } L > 0,\ \mathbb{E}_{\mathrm{test}}[\varepsilon'] = \mathbb{E}_{\mathrm{train}}[\varepsilon] = 0. \tag{17}$$

Roughly speaking, the models in the training and test dataset differ by arbitrary shifts in span(**M**) and a scalar factor in the noise distribution.

Consider the parameter $b^{\to\infty}$ as defined in (7),

$$b^{\to\infty} = \operatorname*{argmin}_{b \in I} \mathbb{E}_{\mathrm{train}}[(Y - X^\intercal b)^2],$$
$$I = \{b; \mathbb{E}_{\mathrm{train}}[Y - X^\intercal b | A] \equiv 0\},$$

which is a functional of the distribution in model (15). For its analogue on a new test dataset with observed variables A', X', Y' we define

$$b'^{\to\infty} = \operatorname*{argmin}_{b \in I'} \mathbb{E}_{\mathrm{test}}[(Y' - (X')^\intercal b)^2],$$
$$I' = \{b;\ \mathbb{E}_{\mathrm{test}}[Y' - (X')^\intercal b | A'] \equiv 0\}.$$

Theorem 2 (Replicability of $b^{\to\infty}$)**.** *Consider the models in* (15) *and* (16) *for the training and test data, respectively. Assume* (17) *and* $\mathbb{E}_{train}[AA^\intercal]$ *and* $\mathbb{E}_{test}[A'(A')^\intercal]$ *are invertible and assume that the projectability condition* (13) *holds.*

Then,

$$b'^{\to\infty} = b^{\to\infty}.$$

Replicability of statistical estimands is arguably a desirable property, but it is a separate question whether $b^{\to\infty}$ is a meaningful quantity. As discussed at the beginning of this section, $b^{\to\infty}$ has an interpretation as a coefficient vector that optimizes a certain worst-case risk. Beyond this interpretation, we believe that the role of A matters to determine whether the components of $b^{\to\infty}$ are scientifically relevant. Loosely speaking, in instrumental variables settings, A induces associations between X and Y that are due to the causal pathway between X and Y. Hence, $b^{\to\infty}$ has a scientific interpretation as the causal effect from X to Y. However, if A plays the role of a confounder (a variable that induces spurious associations between X and Y), then it is common practice to adjust for A, leading to b^0. Under slightly weaker assumptions than in the result above we also get replicability of b^0. In practice, there may be some uncertainty about whether A is an instrument or a confounder, or whether both sets of assumptions are violated. In the next section we will show that anchor regression can be used in such settings to screen for replicable coefficients that have a causal interpretation.

3.2 Anchor stability

If all solutions of *anchor regression* agree (i.e., if $b^0 = b^\gamma$ for all $\gamma \in [0, \infty)$) we call the coefficient vector *anchor stable.*

We will show that under anchor stability we have predictive stability and replicability of variable selection under certain perturbations. Additionally, we will show that *anchor stability* allows a causal interpretation of the coefficient vector under otherwise comparatively weak assumptions. As in the previous section, in the following we assume that the limit $b^{\to\infty} := \lim_{\gamma\to\infty} b^\gamma$ exists.

One of the anchor stability results (Theorem 4) can be generalized to cases where the anchor is endogeneous. This relaxation is relevant for our application in Section 5.1. A rigorous treatment of endogeneous anchors warrants the introduction of a class of models that subsumes acyclic models in Section 2.1. Thus, for reasons of readability we defer the most general version of the theorem to the Appendix, Section 8.13.

Our first result shows that we have anchor stability if the two endpoints of anchor regression agree.

Proposition 1. *If* $b^0 = b^{\to\infty}$ *then*

$$b^0 = b^\gamma \text{ for all } \gamma \in (0, \infty).$$

The proposition is valid without necessarily assuming the projectability condition, which is, however, needed for the following result on anchor stability in the case that the solutions match for $\gamma \in \{0, \infty\}$.

Theorem 3 (Anchor stability, predictive stability and replicability)**.** *Let the assumptions of Section 2.1 hold, and in addition assume the projectability condition* (13) *and that the Gram matrix* $\mathbb{E}_{train}[AA^\intercal]$ *is invertible. If* $b^0 = b^{\to\infty}$*, then, for all random or constant vectors* v *that are uncorrelated of* ε *and take values in span*$(\mathbf{M})$*,*

1. $\mathbb{E}_{train}[(Y - X^\intercal b^0)^2] = \mathbb{E}_v[(Y - X^\intercal b^0)^2]$*, and*
2. $b^0 = \operatorname{argmin}_b \mathbb{E}_v[(Y - X^\intercal b)^2]$.

Part (a) of the theorem implies that the risk is constant as long as the perturbations v lie in the span of the shift matrix $\mathbf{M}$, i.e. in span$(\mathbf{M})$. This can be seen as a form of predictive stability across a range of distributions. Part (b) together with Proposition 1 imply that running a regression on perturbed data sets in the population case returns the same coefficients as the ones computed on the training data as long as the perturbations v lie in span$(\mathbf{M})$. In this sense, we have replicability across certain distributions.

Now let us turn to the interpretation of the individual coefficients in this case. The individual coefficients can be interpreted using the concepts of d-separation, causal directed acyclic graphs and do-interventions. For reasons of readability and as the concepts are otherwise not needed in this paper, we will not define them here but rather refer the reader to e.g. Pearl [2009], Chapter 1. An interpretation of the result in the one-dimensional case is given in Section 3.3. The faithfulness assumption [Spirtes et al., 2000, Pearl, 2009] connects d-separation statements to statements of conditional independences. As *anchor regression* only deals with covariances, we have to make an assumption that connects d-separation statements to partial correlations. We assume that G is acyclic and that for every disjoint sets of variables $V_1, V_2, V_3 \subset (X, Y, H, A)$, V_1 is d-separated of V_2 in G given V_3 if and only if the partial correlation part.cor$(V_1, V_2|V_3) = 0$. This can be seen as a linear version of faithfulness.

Theorem 4 (Anchor stability implies causality)**.** *Let the assumptions of Section 2.1 hold with an acyclic graph* G*, and assume the projectability condition* (13)*.*

Furthermore, assume that for every disjoint sets of variables $V_1, V_2, V_3 \subset (X, Y, H, A)$*,* V_1 *is d-separated of* V_2 *in* G *given* V_3 *if and only if the partial correlation part.cor*$(V_1, V_2|V_3) = 0$*. Furthermore assume that for each* X_k *there exists* k' *such that* $A_{k'} \to X_k$*. If* $b^{\to\infty} = b^0$*, then*

$$b^{\to\infty} = b^0 = \partial_x \mathbb{E}[Y|do(X = x)], \tag{18}$$

where the do-operator $\mathbb{E}[\bullet|do(X = x)]$ *is defined as in Pearl [2009], Chapter 1. In addition, there is no confounder between* X *and* Y*, i.e., there is no* H_k *that is both an ancestor of some* $X_{k'}$ *and* Y *in* G*.*

A more general version of this result that allows for endogeneous anchors can be found in Section 8.13. Roughly speaking, the theorem says that under anchor stability, the coefficients $b^{\to\infty} = b^0$ have a causal interpretation and there is no confounder between X and Y. If confounders were present between X and Y, intervening (or conditioning) on them could potentially change the anchor regression coefficient b^0. In this sense, the absence of confounding between X and Y may be seen as a positive indication for distributional replicability.

Anchor stability is testable on data and if it holds, under relatively weak assumptions, the coefficients allow for a causal interpretation. In empirical studies using instrumental variables, one often compares IV estimates with OLS estimates. The above result formalizes the implications when these estimates are equal.

3.3 Anchor stability in the one-dimensional case

In the special case where X, Y, H and A are all one-dimensional random variables, the theorem can be interpreted in the following way: Suppose we know that A is exogeneous and $A \to X$ but we do not know whether it is a valid instrument, i.e., potentially we have $A \to Y$ or $A \to H \to Y$. We

may not know either whether we could obtain the causal coefficients by simply regressing Y on X or Y on (X, A), i.e., we are unsure whether there exists a hidden confounder H with $X \leftarrow H \rightarrow Y$. Under the assumptions of Theorem 4 and if $b^0 \neq 0$, the models agree if and only if $A \rightarrow X \rightarrow Y$ and if no other arrows (or confounders) are present. Using the theorem, if the two anchor solutions agree, then both the IV and regression adjustment are correct for estimating the causal effect. This approach is restrictive, but can potentially be useful in cases where we have little knowledge about the underlying structure and not much reason to prefer one of these models over the other. An application of this approach is shown in the data section. We anticipate that the concept of *anchor stability* is most useful for screening causal effects in large-scale settings. An analogous statement holds for the multivariate case.

4 Properties of anchor regression estimators

In this section we discuss the properties of finite-sample *anchor regression*. Section 4.1 treats the low-dimensional case; the high-dimensional case is discussed in Section 4.2. In the following we assume to have n i.i.d. observations of (X, Y, A). Concatenating the observations of X row-wise forms an $n \times d$-dimensional matrix that we denote by $\mathbf{X}$. Analogously, the matrix containing the observations of A is denoted by $\mathbf{A} \in \mathbb{R}^{n \times q}$ and the vector containing the observations of Y is denoted by $\mathbf{Y} \in \mathbb{R}^n$. In the following, we tacitly assume that the population parameter b^γ as defined in equation (6) is unique.

4.1 Estimator in the low-dimensional setting

As discussed before, in the low-dimensional case where $d < n$ we recommend using a simple plug-in estimator for the anchor-regression coefficient b^γ:

$$\hat{b}^\gamma = \underset{b}{\operatorname{argmin}} \|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma \|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2, \tag{19}$$

where $\Pi_{\mathbf{A}} \in \mathbb{R}^{n \times n}$ is the matrix that projects on the column space of $\mathbf{A}$, i.e., if $\mathbf{A}^\intercal\mathbf{A}$ is invertible, then $\Pi_{\mathbf{A}} := \mathbf{A}(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal$. In Section 2.1 we made the assumption that X and Y have mean zero. Hence, in practice, we recommend to center $\mathbf{X}$ and $\mathbf{Y}$ in a pre-processing step.

Computation of the *anchor regression* estimator in (19) is simple, as it can be cast as an ordinary least squares problem on a transformed data set. To this end, define

$$\tilde{\mathbf{X}} := (\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{X} + \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{X} \quad \text{and} \quad \tilde{\mathbf{Y}} := (\mathrm{Id} - \Pi_{\mathbf{A}})\mathbf{Y} + \sqrt{\gamma}\Pi_{\mathbf{A}}\mathbf{Y}. \tag{20}$$

The estimator in (19) can then be represented as follows:

$$\hat{b}^\gamma = \underset{b}{\operatorname{argmin}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}b\|_2^2.$$

The transformed data set $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ can be interpreted as artificially generated interventional ("perturbed") data. In this sense, *anchor regression* can be seen as a two-step procedure. First, generate perturbed data $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ for a given perturbation strength γ. Then, run ordinary least squares on the artificial data set.

By the law of large numbers for $n \rightarrow \infty$ the empirical covariance matrix of (X, Y, A) converges to the population covariance matrix of (X, Y, A). By continuity, $\hat{b}^\gamma = (\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{Y}}$ converges to the population parameter b^γ. Hence, $\hat{b}^\gamma$ is a consistent estimator of b^γ.

The transformation (20) is for computational reasons only.

Even if (X, Y, A) follows a multivariate Gaussian distribution, in general it might *not* be true that $\hat{b}^\gamma \sim \mathcal{N}(b^\gamma, V)$ for some covariance matrix V since possible confounding complicates the matter. Hence p-values or confidence intervals from ordinary least squares regression of the transformed data $(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}})$ cannot be used.

Since a main goal in this paper is to establish good predictive performance on future data sets, it is less important to provide distributional results for $\hat{b}^\gamma - b^\gamma$, than to quantify the excess predictive risk on new data sets. A finite sample bound for the excess risk, even covering the high-dimensional setting, can be found in Section 4.3.

4.2 Estimator in the high-dimensional setting

If the number of predictors d exceeds the number of observations n, then the sample estimate defined in (20) is not well-defined. In high-dimensional settings, one typically employs ℓ_1- or ℓ_2-norm penalties for regularization and shrinkage. The ℓ_1-penalized estimators are usually consistent under appropriate sparsity and distributional assumptions, see for example Bühlmann and van de Geer [2011].

While high-dimensionality is allowed in terms of $d \gg n$, we will assume here that the number of anchor variables q is of smaller order than n. High-dimensionality in terms of $q \gg n$ would be another issue, as Π_A is ill-posed, and should be addressed with an ℓ_∞ regularization scheme, replacing the ℓ_2-norm term $\gamma\|\Pi_A(Y - X^\intercal b)\|_2^2$. We propose high-dimensional estimation of *anchor regression* as a solution of

$$\hat{b}^{\gamma,\lambda} = \operatorname*{argmin}_b \|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma\|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2 + 2\lambda\|b\|_1. \tag{21}$$

Compared to unregularized *anchor regression*, the penalty term $2\lambda\|b\|_1$ favours coefficient vectors b that are sparse. For $\gamma = 1$, the estimator coincides with the Lasso [Tibshirani, 1996], whereas for $\lambda = 0$, the estimator coincides with unregularized *anchor regression*.

As in the low-dimensional case with the linear transformation in (20), computation of regularized *anchor regression* is easy. We can rewrite regularized *anchor regression* as

$$\begin{aligned} &\operatorname*{argmin}_b \|(\mathrm{Id} - \Pi_{\mathbf{A}})(\mathbf{Y} - \mathbf{X}b)\|_2^2 + \gamma\|\Pi_{\mathbf{A}}(\mathbf{Y} - \mathbf{X}b)\|_2^2 + 2\lambda\|b\|_1 \\ =&\operatorname*{argmin}_b \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}b\|_2^2 + 2\lambda\|b\|_1, \end{aligned}$$

where $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$ are defined as in equation (20). Hence, solving a high-dimensional *anchor regression* for fixed γ is reduced to solving a Lasso problem. This is typically done by coordinatewise descent [Friedman et al., 2007] to approximately compute the solution path. In the next section we will investigate finite-sample performance of ℓ_1-norm regularized *anchor regression*.

4.3 Finite-sample bound for discrete anchors

We will derive a finite sample bound for discrete anchors. There are no fundamental issues that prevent the derivation of similar results for continuous anchors. We write $\mathcal{A}$ for the set of levels of the random variable A. Unbalanced settings can impose difficulties in the finite-sample case as it becomes more challenging to estimate the penalty term. We analyse the behaviour of *anchor regression* in the case where all anchor levels $A = a$, $a \in \mathcal{A}$, are explicitly given equal weight in the optimization procedure, i.e., the objective function for population *anchor regression* is

$$R(b) := \mathbb{E}_{\text{train}}[(Y - X^\intercal b - \mathbb{E}_{\text{train}}[Y - X^\intercal b|A])^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mathbb{E}_{\text{train}}[Y - X^\intercal b|A = a])^2.$$

Such a re-weighting is usually advisable in unbalanced settings. Otherwise, very few levels of A can dominate the penalty term and limit its usefulness. Note that, by Theorem 1, $R(b)$ corresponds to the maximum ℓ_2-risk under a uniform distribution on the levels of A:

$$R(b) = \sup_{v\in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

For data with unbalanced discrete anchor levels, the shape of C^γ changes as anchor levels that occur with small probability are given less weight. For discrete anchors, interpreting *anchor regression* via quantiles is only justified under re-weighting, see Lemma 3 in the Appendix..

To formulate the assumptions in a convenient form we introduce additional notation for the special case of discrete anchors. We write n_a for the number of observations for level $A = a$ and $n_{\min}$ for the minimum number of observations, i.e., $n_{\min} := \min_{a\in\mathcal{A}} n_a$. We write $\mathbf{X}^{(a)} \in \mathbb{R}^{n_a\times d}$ for the observations for which $A = a$. In other words, the rows of $\mathbf{X}^{(a)}$ consist of observations $\mathbf{X}_{i,\bullet}$ for which $\mathbf{A}_i = a$. Furthermore we write $\overline{\mathbf{X}}^{(a)}$ for the mean within the group, i.e., $\overline{\mathbf{X}}^{(a)} = \frac{1}{n_a}\sum_{i=1}^{n_a}\mathbf{X}^{(a)}_{i,\bullet}$. Analogously we define $\mathbf{Y}^{(a)} \in \mathbb{R}^{n_a}$ and $\overline{\mathbf{Y}}^{(a)}$. Using this notation, the high-dimensional *anchor regression* estimator in (21) but with equal weight regularization, analogous to the definition of $R(b)$ above, equals

$$\hat{b} := \operatorname*{argmin}_b \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left(\mathbf{Y}^{(a)}_i - \overline{\mathbf{Y}}^{(a)} - (\mathbf{X}^{(a)}_{i,\bullet} - \overline{\mathbf{X}}^{(a)})b\right)^2 + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\left(\overline{\mathbf{Y}}^{(a)} - \overline{\mathbf{X}}^{(a)}b\right)^2 + 2\lambda\|b\|_1.$$

Here and in the following, we suppress the dependence of $\hat{b}$ on γ and λ. For any $S \subseteq \{1, \dots, d\}$ and stretch factor $L > 0$ define the *anchor compatibility constant*

$$\hat{\phi}^2(L, S) := \min_{\|b_S\|_1 = 1, \|b_{-S}\|_1 \leq L} |S| \left(\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} \left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)}) b \right)^2 + \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\overline{\mathbf{X}}^{(a)} b)^2 \right).$$

To proceed, we need a lower bound on the compatibility constant $\hat{\phi}^2(L, S^*)$ for $S^* := \{k : b_k^\gamma \neq 0\}$, the active set of b^γ. Note that for all S

$$\hat{\phi}^2(L, S) \geq \min(\gamma, 1) \min_{\|b_S\|_1 = 1, \|b_{-S}\|_1 \leq L} \frac{|S|}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} \left(\mathbf{X}_{i,\bullet}^{(a)} b \right)^2.$$

For $|\mathcal{A}| = 1$ the quantity on the right corresponds to the ordinary compatibility constant in high-dimensional linear regression [van de Geer, 2016]. The anchor compatibility constant can be bounded analogously as the ordinary compatibility constant, see e.g. van de Geer [2016].

When presenting asymptotic results as both $d = d_n > n \to \infty$, we allow that the set $\mathcal{A}$, the shift matrix $\mathbf{M}$, the target quantity b^γ and the structural equation model change for varying n.

Theorem 5. *Consider the model in* (8) *and assume that ε is multivariate Gaussian. Moreover, assume that $(\mathbf{X}_{i,\bullet}^{(a)}, \mathbf{Y}_i^{(a)})$, $i = 1, \dots, n_a$, are i.i.d. random variables that follow the distribution of $(X, Y)|A = a$ under $\mathbb{P}_{train}$. Fix $\gamma > 0$ and assume that $\hat{\phi}^2(8, S^*) \geq c$ for some constant $c > 0$ with probability $1 - \delta$, and that $S^* \neq \emptyset$. Choose $t \geq 0$ such that*

$$|S^*|^2 (t + \log(d) + \log(|\mathcal{A}|))/n_{\min} \leq c',$$

for some constant $c' > 0$. Then, for $\lambda \geq C\sqrt{(t + \log(d) + \log(|\mathcal{A}|))/n_{\min}}$, with probability exceeding $1 - 10\exp(-t) - \delta$,

$$R(\hat{b}) \leq \min_b R(b) + C' \lambda^2 |S^*|,$$

where the constants $C, C' < \infty$ depend on $\max_k \mathrm{Var}(X_k)$, $\mathrm{Var}(Y - X^\intercal b^\gamma))$, $\max_{a \in \mathcal{A}} \|\mathbb{E}_{train}[X|A = a]\|_\infty$, $\max_{a \in \mathcal{A}} |\mathbb{E}_{train}[Y - X^\intercal b^\gamma | A = a]|)$, γ, c and c'. The variances are meant with respect to the measure $\mathbb{P}_{train}$.

There are no fundamental issues that prevent the derivation of similar results for continuous anchors. The constant 8 in the anchor compatibility constant $\hat{\phi}^2(8, S^*)$ does not represent a theoretically meaningful critical value, it was chosen in an ad-hoc fashion to simplify the result.

Under the assumptions mentioned above, if we choose $\lambda \asymp \kappa C\sqrt{(t + \log(d) + \log(|\mathcal{A}|))/n_{\min}}$ for $\kappa > \sqrt{2}$, $t = \log(d)$ and assume that $\delta \to 0$, we obtain the following asymptotic result. For $d, n \to \infty$, with probability going to one,

$$R(\hat{b}) - \min_b R(b) = \mathcal{O}\left(\frac{|S^*|(\log(d) + \log(|\mathcal{A}|))}{n_{\min}} \right).$$

As $\hat{b}$ coincides with the Lasso for $\gamma = 1$ and $|\mathcal{A}| = 1$, it is worthwhile to compare this bound to risk bounds of the Lasso. The excess predictive risk of the Lasso in a comparable setting with appropriate choice of λ is of the order $\mathcal{O}(|S^*| \log(d)/n)$, see, e.g., Bühlmann and van de Geer [2011, Chapter 6]. Hence the risk bounds will be of comparable order as long as $n/n_{\min}$ is bounded.

5 Numerical examples

We provide two numerical examples. The first example shows how anchor regression can be used to improve replicability across perturbed data. In the second example, we discuss a prediction problem under distributional shifts. The code is available on `github.com/rothenhaeusler`.

5.1 Genotype-tissue expression

The data was obtained from the Genotype-Tissue Expression (GTEx) portal [Carithers et al., 2015]. One of the GTEx datasets contains gene expression data from 53 tissues of 714 human donors, in total comprising $n = 11688$ observations of $d = 12948$ genes.

These samples were collected postmortem. Gene expressions are subject to various types of heterogeneity. They vary not only between humans but also between different tissues and individual cells. 13 out of the 53 tissues contain more than 300 observations. We conducted our analysis on these 13 tissues.

We will compare features that are relevant for prediction on one tissue with the features that are relevant for prediction on another tissue. Our goal is to find relevant features that are not particular to the specific tissue at hand, but can also be found (replicated) on the other tissues. Due to the heterogeneity between the tissues, this is a challenging task. The response variable Y is the expression of a target gene and the covariates X are the expressions of all other genes. Mathematically, we associate with $y \in \{1, \ldots, d\}$ the gene index of the target variable and $x = \{1, \ldots, d\} \setminus y$ the gene indices of the expression covariates.

For each tissue, the gene expressions and additional covariates are available. These covariates contain geno-typing principal components, PEER factors, sex and genotyping platform. The geno-typing principal components and PEER factors (which are constructed from covariates and gene expressions) account for some (but not all) of the confounding sources of expression variation, such as batch effects, environmental influences and sample history [Stegle et al., 2012]. Originally, it has been suggested to include the PEER factors when regressing gene expression on genotype. Here, we use them in an analysis of co-expression, in spirit similarly to Furlotte et al. [2011] or Stegle et al. [2011]. We will use these additional covariates as the anchor variables[1]. We consider combinations of biological entities, and the PEER factors are partially computed from the gene expressions. Therefore, strictly speaking, the assumptions in Section 3.2 are not satisfied. Assuming, however, that these PEER factors and geno-typing principal components are correlated with confounding sources of variation, using anchor stability with these proxy variables as anchor may still increase replicability of feature selections across data sets. Note that using anchor stability is justified even in cases where anchors are endogeneous, see the discussion in Section 3.2 and the corresponding theorem in the Appendix, Section 8.13.

5.1.1 Improved replicability with stable anchor regression

The goal is to investigate whether features that are relevant for prediction on one tissue are also relevant for prediction on other tissues. More specifically, we compute and rank variables using the Lasso and penalized *anchor regression* on one specific tissue t. Then, we check whether the discoveries can also be replicated on the other tissues $t' \neq t$.

How should we rank the covariates in an *anchor regression* framework? By the discussion below Theorem 4, anchor stability is potentially a positive indicator for distributional replicability. This suggests that ranking by anchor stability should improve replicability across heterogeneous domains of the data set. In cases where the anchor is only weakly correlated with the covariates, estimation of b^γ will be unstable for $\gamma \to \infty$. Thus, in the following, we do not test whether the coefficients are invariant across $\gamma \in [0, \infty)$ but check whether the individual *anchor regression* coefficients are bounded away from 0 for $\gamma \in [0, 1]$. This can be seen as a weak form of anchor stability.

Consider a fixed tissue t. For the *anchor regression* method, we compute

$$a_{y,k,t} := \min_{\gamma \in [0,1]} |\hat{b}_k^{\gamma,\lambda}|, \tag{22}$$

where $\hat{b}^{\gamma,\lambda}$ is the $p-1$-dimensional anchor coefficient of a *anchor regression* of target variable $y \in \{1, \ldots, p\}$ on the other gene expressions $x = \{1, \ldots, p\} \setminus \{y\}$. As regularization parameter λ we use the same as for the Lasso regression (see below). We also consider for (22) the ranges $\gamma \in \{[0, 0.25], [0, 16]\}$ and show the results in Section 8.17.

For comparison, we compute the Lasso coefficients

$$l_{y,k,t} := |(\hat{b}_{\text{lasso}})_k|, \tag{23}$$

where $\hat{b}_{\text{lasso}}$ is the $p-1$-dimensional Lasso coefficient of a Lasso regression of target variable $y \in \{1, \ldots, d\}$ on all other variables $x = \{1, \ldots, d\} \setminus y$, after removing the effect of the anchor variables. By definition, $\hat{b}_{\text{lasso}} = \hat{b}^{0,\lambda}$, i.e. the Lasso coefficient vector coincides with *anchor regression* for $\gamma = 0$ which implies $a_{y,k,t} \leq l_{y,k,t}$. Hence, any nonzero effect found using anchor regression is also

[1] From a theoretical standpoint, using the tissues as anchor is a reasonable choice as well. However, the empirical conditional expectations of each gene expression given the tissues is zero. The gene expressions have been normalized within each tissue and hence using the tissues as the anchor variable is not meaningful for this dataset.

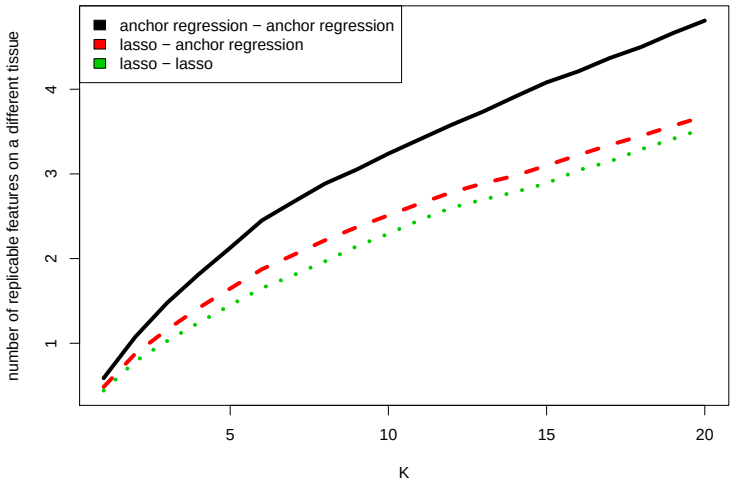


Figure 4: Replicability of variable selection in GTEx data. Plotting how many of the $K \in \{1, \ldots, 20\}$ top-ranked features found by *anchor regression* and Lasso on one tissue t are also one of the K top-ranked features on another tissue t'. The results are summed over all other tissues $t' \neq t$, averaged over all tissues t and averaged over 200 random choices of y, and they are plotted as y-coordinates. For *anchor regression* the ranking is according to (22), and for Lasso according to (23). The legend describes the method used on one tissue t and the method used on another tissue t'. *Anchor regression* exhibits the highest degree of replicability.

a nonzero effect using the Lasso. However the ranking for the two methods is different. For both methods, a regularization parameter λ has to be chosen. We use the one from cross-validation as implemented in the function `cv.glmnet` in the `R`-package `glmnet`. To make the methods comparable, this regularization parameter was also used for the *anchor regression* method.

We evaluate how many of the largest effects found by stable anchor regression or Lasso can be replicated on another tissue. The results are depicted in Figure 4. The black solid line depicts how many of the $K = 1, \ldots, 20$ largest effects $l_{y,k,t}$ are also among the K largest effects $l_{y,k,t'}$ on another tissue $t' \neq t$ for a fixed target y (and then averaged over y, see below). Analogously, the red dashed line shows how many of the K largest effects $a_{y,k,t}$ are also among the K largest effects $l_{y,k,t'}$ on a tissue $t' \neq t$. Finally, the green dotted line shows how many of the K largest effects $a_{y,k,t}$ are also among the K largest effects $a_{y,k,t'}$ on a tissue $t' \neq t$. The results are summed over all choices of $t' \neq t$ and averaged over 200 random choices of $y \in \{1, \ldots, 12948\}$.

Both anchor stable and Lasso methods are better than random guessing. Ranking by anchor stable regression results in improved replicability across tissues. Note that this is a challenging data set and the predictive power among genes is small: the average R^2 for a Lasso run estimated and evaluated on disjoint parts of one tissue is .37. The average R^2 for a Lasso run estimated on one tissue and evaluated on another tissue is slightly negative. In Section 8.17, we also discuss the degree of replicability for the parameter $b^{\to\infty}$.

5.2 Bike sharing data set

The data set is taken from the UCI machine learning repository [Fanaee-T and Gama, 2013, Dheeru and Karra Taniskidou, 2017]. It contains $n = 17379$ hourly counts of bike rentals from 2011 to 2012 of the Capital bike share in Washington D.C. The goal is to predict bike rentals (variable cnt) using weather data reliably across days. As the variable cnt is a count, a square-root transformation was carried out. The effect of categorical variables, for which shift interventions are not meaningful (this includes the variables working day, weekday, holiday), was removed in a pre-processing step. While we generally recommend removing the effect of variables that cannot be shifted, in this particular example the pre-processing step makes no discernible difference in the resulting plot, see Figure 12

in the Appendix. The data set contains the numerical covariates temperature, feeling temperature, humidity and windspeed. The variable hour is nested within the variable "date". We will first conduct the analysis ignoring the variable "hour" as this application is closest to Theorem 1, Lemma 1 and Lemma 3. In practice, one would also want to include "hour" as a predictor in the model. We discuss this case further below.

There are large fluctuations in the usage of bikes that cannot be explained by weather data alone [Fanaee-T and Gama, 2013]. Instead of using the discrete variable 'date' for prediction, we use it as an anchor A. More detailed, the anchor variable is discrete with one level per day.

This choice of anchor variable allows us to investigate the performance of the algorithm in a setting with strong heterogeneities. The goal is to predict the count of bike rentals in a reliable fashion using the covariates temperature, feeling temperature, humidity and windspeed.

As evaluation metric, we consider quantiles of the conditional mean squared error given the anchor variable. Intuitively speaking, we want to train a prediction rule that works reliably across days. Practically, this means that for each fixed day, we average over the prediction loss and then compute quantiles across days. The quantiles of the conditional squared error $\mathbb{E}[(Y - X^\intercal b)^2|A]$ are a proxy for the right-hand side of equation (11) being the worst case risk across perturbations of a certain level, cf. Lemma 3 in the Appendix. The data was split into 5 consecutive blocks. The estimator was trained on 4 of the 5 blocks and tested on the left-out block. Results are averaged over the five possible train-test split. Quantiles of the daily averaged squared error on the test data set $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$, are depicted in Figure 5.

The optimal choice of γ as evaluated on the test data set as a function of the quantile and the corresponding predictive performance can be found in Figure 6. This motivates choosing γ by minimizing quantiles of the loss on held-out data. We describe this procedure in more detail below. Figure 5 shows that for small quantiles, small values of γ are slightly preferred, while for quantiles close to one, large values of γ clearly outperform smaller values. This is in line with the theory presented in Section 2.4.

However, as the direction and strength of the perturbations usually also changes to some extent between training and test data set we do not recommend simply using $\lim_{\gamma\to\infty} \hat{b}^\gamma$. In practice, we do not advise to choose γ based on Lemma 1 or Lemma 3 as the interplay of the penalization parameter and quantiles of $\mathbb{E}[(Y - X^\intercal b^\gamma)^2|A]$ is more involved for non-Gaussian distributions. Instead, we recommend choosing an optimal γ based on cross-validation.

The cross-validation approach (as used in Figure 6) proceeds as follows. First, choose a quantile α (for example $\alpha = 90\%$). In each of the folds, the data is split in a training data set and a test data set, such that each level of the anchor variable only appears in one of the data sets. Then, for varying γ, compute $\hat{b}^\gamma$ on the training data set and estimate the α-quantile of $\mathbb{E}[(Y - X^\intercal b^\gamma)^2|A]$ on the test data set. After averaging the estimated quantiles over the folds, choose γ such that the chosen quantile is minimized. For this approach to work, we have to make an assumption that heterogeneities of the future data generating process are in some sense similar to the heterogeneities observed in the training data set. This assumption is made precise in Lemma 3 in the Appendix for discrete anchors.

As discussed above, the application above is close to the theory presented in Section 2, but in practice one would also want to include the predictor "hour". As an alternative experiment to the one shown above, we run a regression of the target variable on the predictor "hour" and run anchor regression on the residuals. For the final prediction, we then add the predictions from both models. The variable hour differs from the other variables in the sense that it is nested within the anchor date. Thus, building the overall model in such a hierarchical fashion is not supported by our current theory. The results can be found in the Appendix: Figure 13 in Section 8.18 is equivalent to Figure 5, but *anchor regression* is run on the residuals after regressing out the effect of "hour". For large quantiles of the conditional loss, $\gamma \gg 1$ outperforms $\gamma < 1$, but the relationship is not monotonous. Figure 14 in Section 8.18 of the Appendix is similar to Figure 5 but with the modified anchor regression procedure described above. The anchor regression procedure performs better than ordinary least-squares ($\gamma = 1$) for all considered quantiles.

6 Practical guidance

In this section we summarize our results and give high-level guidance for using *anchor regression*, based on our empirical experience and theoretical results.

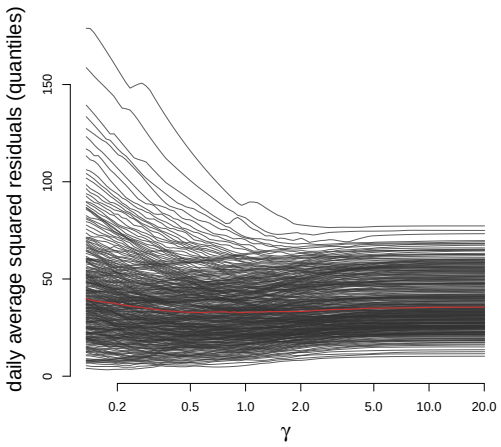


Figure 5: Daily average squared residuals $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$ as a function of γ. Each line corresponds to a quantile of $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$. The quantiles are chosen in the set $\{0.05, 0.01, \ldots, 0.995\}$, with the median marked in red. For growing γ, the upper percentiles of $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$ are decreasing while the lower percentiles are slightly increasing. This is in line with the theory presented in Section 2.4. The distribution of bike rentals is expected to change from day to day. For growing γ, the upper percentiles of the loss are reduced, i.e., predictions are increasingly reliable across days. A comparison to OLS with $\gamma = 1$ is given in the right panel of Figure 6.

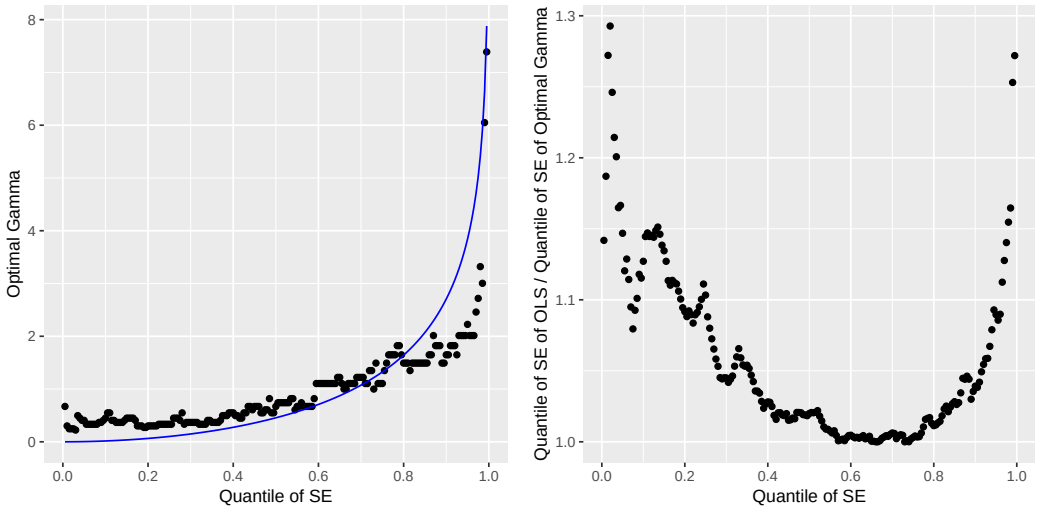


Figure 6: Optimal choice of γ and predictive performance of *anchor regression* for varying quantiles of the squared error on the bike-sharing data set. On the left-hand side, the optimal choice of γ is depicted as a function of quantiles of the daily averaged error, $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$. The blue line shows the theoretically optimal choice of γ using Lemma 1. The black dots show the optimal choice of γ as evaluated on the test data set. For growing quantiles, the optimal choice $\gamma = \gamma_{\text{opt}}$ increases. For example, $\gamma \approx 0.35$ is optimal for minimizing the 5%-Quantile of $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$. Similarly, $\gamma \approx 2$ is optimal for minimizing the 90%-Quantile of $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$. On the right-hand side, the performance with the optimal estimated γ is shown in terms of quantiles of $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$, relative to ordinary least squares (OLS). For example, for the 90%-quantile, the optimal choice of γ leads to a 10%-improvement of *anchor regression* compared to ordinary least squares. The biggest improvements compared to OLS are obtained for both very small and very large quantiles. The quantiles of $\hat{\mathbb{E}}_{\text{test}}[(Y - X^\intercal \hat{b}^\gamma)^2|A]$ were estimated using 5-fold cross-validation.

Possible Applications. *Anchor regression* can be applied in settings, where we are given data from a target variable Y and covariates X and are interested in generalizing across heterogeneous data sets. Examples of such distribution changes include batch effects, population shifts, and heterogeneity across time or locations. In the case of prediction, the approach aims to achieve robust predictions across data sets. *Anchor regression* is optimal if the data sets differ by (restricted) shift interventions. For the goal of parameter estimation, *anchor regression* can be used to find features that are invariant across a (restricted) set of distributions, see Section 3.2. Thus, the approach might help to increase the replicability of discoveries across data sets.

Choice of the anchor variable. In the case of prediction, the main assumptions are linearity of the system and exogeneity of the anchor. We recommend to choose the anchor based on the type of robustness or invariance one aims to obtain. For example, if one intends to obtain robustness of the prediction rule across locations, we recommend using location as an anchor variable. If the goal is to achieve robustness across time, we recommend using discretized time windows as an anchor variable. In our theory, this recommendation is justified by Theorem 1. Different choices of the anchor correspond to different matrices $\mathbf{M}$, which in turn provide protection against different distributional shifts.

In the case of estimation, the exogeneity assumption for the anchor variable can be dropped. Details can be found in Section 3.2 and in Section 8.13 in the Appendix. In that case, the anchor should be chosen such that it affects the covariates of interest as much as possible.

Choice of the regularization parameter. When using *anchor regression* for prediction, one has to choose a regularization parameter γ. If possible, this should be done based on subject matter knowledge. For example, if one expects perturbations on future data sets to be at most 1.5 times as large as on the training data sets, $\gamma = 1.5$ is a sensible choice. If the anchor variable has many categorical levels, it is also possible to choose γ using some form of leave level out cross-validation. This approach is described in Section 5.2. For data sets where the above considerations do not apply, we believe that $\gamma = 2$ is a good default choice.

For screening via anchor stability, in theory it is sufficient to test whether the two endpoints $\gamma = 0$ and $\gamma = \infty$ of *anchor regression* agree, see Proposition 1. In cases where the anchor is only weakly associated with the covariates, estimation of b^∞ will be unstable. Thus, in practice we recommend to screen based on a weak form of anchor stability, as in equation (22). That choice can be considered a heuristic, as its theoretical implications are yet to be investigated.

Limitations. All extrapolation statements of *anchor regression* rely on the assumption of linearity. Using *anchor regression* for prediction generally does not guarantee protection against "black swan events". More specifically, *anchor regression* is not leading to robust prediction when the heterogeneity between the data sets is different from the restricted set of shift interventions that have been observed on the training data sets.

For example, in Theorem 1, the set C^γ contains shifts that lie in the span of $\mathbf{M}$, as opposed to shifts in arbitrary directions. In cases where distribution shifts are complex, in the sense that distributions change arbitrarily between data sets, neither *anchor regression* nor any other method can provide reliable predictions. If the anchor does not shift any distributions, i.e. if the distribution of (X, Y) is constant across values of A then there is no benefit from using the *anchor regression* approach. Note however, that in this case there is also little harm from using the *anchor regression* approach as the penalty term in equation (11) will be close to zero.

7 Discussion and outlook

We have introduced *anchor regression*, a regularization approach for fitting linear models. We have shown that this approach optimizes worst case prediction risk over a class of perturbations and that it also leads to improved replicability of variable selection across different perturbed heterogeneous datasets. The methodology has relations to invariance properties from causality and the concrete proposed procedure of *anchor regression* interpolates between three common statistical estimation schemes, namely partialling out (i.e., adjusting for) exogeneous variables, ordinary least squares and two-stage least squares from instrumental variables regression (with exogeneous instruments).

The penalty in *anchor regression* corresponds to the change in prediction loss under certain perturbations. More specifically, these perturbations are modelled as random or deterministic shift interventions and are estimated from a heterogeneous training data set. We have explored the prediction behavior, both in terms of size and direction of the considered perturbations. When considering the regularization path of *anchor regression* as a function of the penalty or regularization parameter, we prove some stability and replicability for variable importance or variable selection over a range of perturbations, i.e., a range of potentially new heterogeneous data sets. Thus, *anchor regression* also contributes to much desired improved replicability of variable importance. We also derived a finite sample bound for worst case prediction in the high-dimensional case.

We consider the behavior of *anchor regression* on real-data applications, in terms of replicability of variable selection and prediction on new potentially perturbed data. We believe that it is worthwhile to explore penalization schemes that exploit heterogeneities that occur in the training distribution and lead to robustness and replicability on new perturbed test data, i.e., generalizing to new unobserved heterogeneity. Such a regularization allows to explicitly balance the tradeoffs between predictive performance on perturbed and unperturbed data sets, while avoiding the loss in prediction accuracy that is incurred when using more conservative approaches (e.g., causal parameters).

Looking ahead, there are some avenues which we think are worthwhile to pursue. In the following, we outline two directions that seem particularly promising.

Beyond shift interventions. Instead of considering shift interventions, it may be interesting to look at penalty schemes that arise from other types of perturbations, such as noise, edge functions and do-interventions. Depending on the application, such interventions may be more appropriate than shift-interventions. In this light, structural equation modelling can serve as a scheme to generate and explore new types of perturbation penalties. Furthermore, it allows to obtain optimality statements to better understand the tradeoffs between perturbation stability and predictive performance.

Nonlinear models. For the *anchor regression* method to be practical in a wide range of scenarios, it is important to extend it to non-linear models. Using a bias-variance decomposition, with $P_A = \mathbb{E}_{\text{train}}[\bullet|A]$ the prediction loss of a non-linear function $g(X)$ can be decomposed as

$$\mathbb{E}_{\text{train}}[(Y - g(X))^2|A] = \mathbb{E}_{\text{train}}[((\text{Id} - P_A)(Y - g(X)))^2|A] + (P_A(Y - g(X)))^2$$

If the conditional variance is constant across strata defined by $A = a$, then the conditional loss simplifies to

$$\mathbb{E}_{\text{train}}[(Y - g(X))^2|A] = \mathbb{E}_{\text{train}}[((\text{Id} - P_A)(Y - g(X)))^2] + (P_A(Y - g(X)|A))^2.$$

This decomposition motivates non-linear *anchor regression*, which we define as the solution to

$$g^\gamma := \arg\min_{g \in \mathcal{G}} \mathbb{E}_{\text{train}}[((\text{Id} - P_A)(Y - g(X)))^2] + \gamma \mathbb{E}_{\text{train}}[(P_A(Y - g(X)))^2],$$

for an appropriate set of functions $\mathcal{G}$. Qualitatively this estimator behaves similarly to *anchor regression*. As before, it interpolates between nonlinear versions of PA, OLS and IV. For $\gamma \to \infty$, non-linear *anchor regression* will strive for invariance in the sense that it tries to keep $\mathbb{E}[(Y - g(X))^2|A]$ constant across all levels of A. The set of interventions that nonlinear *anchor regression* protects against for a fixed γ is not as straightforward to describe as in Theorem 1. However, we conjecture that this estimator behaves similarly to linear *anchor regression* on data sets, in the sense that it potentially improves replicability across heterogeneous regimes and improves robustness of prediction rules across the strata defined by A. Other non-linear extensions of *anchor regression* and some preliminary empirical evidence can be found in Bühlmann [2018]. We believe that it is a promising avenue to further investigate the behaviour of these and related estimators both in theory and practice.

Acknowledgements

We thank Martin Emil Jakobsen for pointing out the link between anchor regression and k-class estimators. We thank several reviewers for various helpful comments. DR received funding from the ONR grant N00014-17-1-2176. PB received funding from the European Research Council under the grant agreement No. 786461 (CausalStats – ERC-2017-ADG).

References

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

K.A. Bollen. *Structural Equations with latent variables*. John Wiley & Sons, 1989.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

R.J. Bowden and D.A. Turkington. *Instrumental variables*, volume 8. Cambridge University Press, 1990.

P. Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: Methods, theory and applications*. Springer, 2011.

L. Carithers, K. Ardlie, M. Barcus, P. Branton, A. Britton, S. Buia, C. Compton, D. DeLuca, J. Peter-Demchok, E. Gelfand, P. Guan, G. Korzeniewski, N. Lockhart, C. Rabiner, A. Rao, K. Robinson, N. Roche, S. Sawyer, A. Segr, C. Shive, A. Smith, L. Sobin, A. Undale, K. Valentino, J. Vaught, T. Young, and H. Moore. A novel approach to high-quality postmortem tissue procurement: The gtex project. *Biopreservation and Biobanking*, 13(5):311319, 2015.

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

D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

V. Didelez, S. Meng, and N.A. Sheehan. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25:22–40, 2010.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74:981–995, 2007.

D. Entner, P. Hoyer, and P. Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial Intelligence and Statistics*, pages 256–264, 2013.

J. Fan and W. Zhang. Statistical estimation in varying coefficient models. *Annals of Statistics*, 27: 1491–1518, 1999.

H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15, 2013. ISSN 2192-6352. doi: 10.1007/s13748-013-0040-3. URL `[WebLink]`.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

W. Fuller. *Measurement error models*, volume 305. John Wiley & Sons, 2009.

N. A. Furlotte, H. M. Kang, C. Ye, and E. Eskin. Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. *Bioinformatics*, 27(13):i288–i294, 2011.

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

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

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55:757–796, 1993.

C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2018.

P.J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

P.J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *Annals of Statistics*, pages 799–821, 1973.

S. Klepper and E. Leamer. Consistent sets of estimates for regressions with errors in all variables. *Econometrica*, pages 163–183, 1984.

K. Korb, L. Hope, A. Nicholson, and K. Axnick. Varieties of causal intervention. In *Proceedings of the Pacific Rim Conference on AI*, pages 322–331, 2004.

S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50: 157–224, 1988.

E. Leamer. Least-squares versus instrumental variables estimation in a simple errors in variables model. *Econometrica*, pages 961–968, 1978.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10846–10856. Curran Associates, Inc., 2018.

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

N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *Annals of Statistics*, 43(4):1801–1830, 2015.

A. Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, pages 575–595, 1959.

S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2nd edition, 2009.

J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, pages 579–595, 2014.

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

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

N. Pfister, S. Bauer, and J. Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.

J.C. Pinheiro and D.M. Bates. Linear mixed-effects models: Basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000.

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

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

D.B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322–331, 2005.

A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *Sixth International Conference on Learning Representations (ICLR)*, 2018.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, 2nd edition, 2000.

O. Stegle, C. Lippert, J. M. Mooij, N. D. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate gaussian models with iid observation noise. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 630–638. Curran Associates, Inc., 2011.

O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

H. Theil. *Economic forecasts and policy*. North-Holland, 1958.

J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 512–522, 2001.

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

S. van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

P.G. Wright. *The tariff on animal and vegetable oils*. The Macmillan company New York, 1928.

H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.

B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.

8 Appendix

8.1 Interpretation of the model class in the cyclic case

If the graph G is cyclic, then the model class in Section 2.1 describes the distribution in an equilibrium state. To see this, let us write

$$V_0 = \varepsilon$$

and

$$V_t = \mathbf{B}V_{t-1} + \varepsilon \text{ for all } t \geq 1,$$

where $V = (X, Y, H)^\intercal$. If the spectral norm of $\mathbf{B}$ is strictly smaller than one, then for each ε we have

$$\lim_{t\to\infty} V_t = \sum_{k\geq 0} \mathbf{B}^k \varepsilon = (\mathrm{Id} - \mathbf{B})^{-1}\varepsilon.$$

Note that if $\mathbf{B}$ is acyclic then this limit always exists. Analogously one can define the shifted distribution as

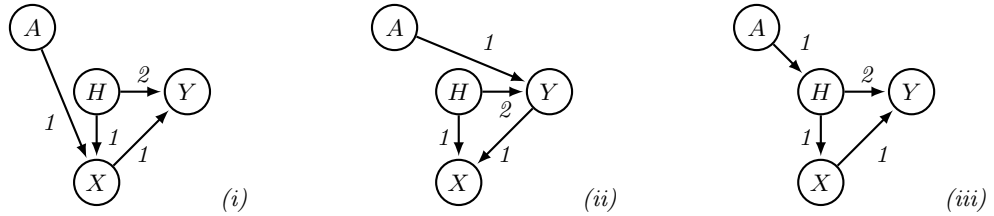
$$\begin{aligned} V_0 &= \varepsilon + v \\ V_t &= (\mathrm{Id} + \mathbf{B})V_{t-1} \\ \lim_{t\to\infty} V_t &= \sum_{k\geq 0} \mathbf{B}^k(\varepsilon + v) = (\mathrm{Id} - \mathbf{B})^{-1}(\varepsilon + v) \end{aligned}$$

ence, by the definition of V, we have $V = \lim_{t\to\infty} V_t$ and our model describes the distribution of a cyclic causal model in its equilibrium.

8.2 Sets C^γ for three examples

In this section we discuss three examples to shed more light on Theorem 1 and the behaviour of *anchor regression*. In particular, the sets C^γ are discussed for the three simple examples. We will see that C^γ can contain interventions not only on X but potentially also on Y and H. The SEM and graph in each case are given in Example 3.

Example 3 (Three SEMs and corresponding sets C^γ)**.** *In each of these SEMs for simplicity we assume that* $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_3)$ *and* $A \sim \mathcal{N}(0, 1)$*. For (i), the corresponding SEM is* $H \leftarrow \varepsilon_3$, $X \leftarrow H + A + \varepsilon_1$, $Y \leftarrow 2H + X + \varepsilon_2$*. In this example,* C^γ *contains interventions on* X *up to strength* γ*, i.e.,* $C^\gamma = \{(t, 0, 0)^\intercal : t^2 \leq \gamma\}$*. For (ii), the corresponding SEM is* $H \leftarrow \varepsilon_3$, $X \leftarrow H + Y + \varepsilon_1$, $Y \leftarrow A + 2H + \varepsilon_2$*. In this example,* C^γ *contains interventions on* Y *up to strength* γ*, i.e.,* $C^\gamma = \{(0, t, 0)^\intercal : t^2 \leq \gamma\}$*. For (iii), the corresponding SEM is* $H \leftarrow A+\varepsilon_3$, $X \leftarrow H+\varepsilon_1$, $Y \leftarrow 2H+X+\varepsilon_2$*. In this example,* C^γ *contains interventions on* H *up to strength* $\sqrt{\gamma}$*, i.e.,* $C^\gamma = \{(0, 0, t)^\intercal : t^2 \leq \gamma\}$*.* C^γ *takes more complex forms when* A *points to several variables. Examples of this phenomenon are discussed in Section 8.4.*



$C^\gamma = \{(t, 0, 0)^\intercal : t^2 \leq \gamma\}$ $C^\gamma = \{(0, t, 0)^\intercal : t^2 \leq \gamma\}$ $C^\gamma = \{(0, 0, t)^\intercal : t^2 \leq \gamma\}$

Example (i) corresponds to a classic IV setting. Here, we have $\mathbf{M} = (1, 0, 0)^\intercal$. Hence, C^γ is the set of interventions on X up to "strength" γ, i.e., $C^\gamma = \{(t, 0, 0)^\intercal : t^2 \leq \gamma\}$. By Theorem 1, b^γ minimizes the ℓ_2-loss under shift interventions on X up to "strength" γ. Similarly for example (ii): *anchor regression* minimizes the ℓ_2-loss under interventions on Y. In example (iii), *anchor regression* minimizes the ℓ_2-loss under interventions on H. In the following we want to investigate whether *anchor regression* can achieve predictive stability, i.e., stable predictive performance in these SEMs under strong interventions. This question can be answered by investigating the limit

$b^{\to\infty} = \lim_{\gamma\to\infty} b^\gamma$. In example (i), we obtain $b^{\to\infty} = 1$. In example (ii), we have $b^{\to\infty} = 1$ and in example (iii) we have $b^{\to\infty} = 3$. A short calculation shows that the distribution of $Y - X^\intercal b$ under $\mathbb{P}_v$ is invariant under shift interventions on X. Formally,

$$Y - X^\intercal b^{\to\infty} \text{ under } \mathbb{P}_v \text{ has the same distribution for all } v = (t, 0, 0)^\intercal.$$

In particular, the MSE $\mathbb{E}_v[(Y - X^\intercal b)^2]$ is constant under shift interventions on X. Similarly in example (ii), the distribution of $Y - X^\intercal b^{\to\infty}$ is invariant under shift interventions on Y. And in example (iii), the distribution of $Y - X^\intercal b^{\to\infty}$ is invariant under shift interventions on H. This holds for any set of edge coefficients with one of the graph structures as in Example 3. However, for some graphs (for example for the graph that arises from reversing the edge between X and Y in (ii)), the invariance statement above does not hold.

In all examples, A is correlated with X. Let c_x denote the effect of A on X, i.e. the regression coefficient when regressing X on A. Let c_y denote the effect of A on Y. Thus, for $b = \frac{c_y}{c_x}$, the effect of A on the synthetic variable $r = Y - Xb$ is zero and r has invariant distribution under conditioning on A. Conditioning on A can be interpreted as certain shift interventions on (X, Y, H). This in turn implies the invariance properties discussed above. Thus, as long as the effect of a one-dimensional anchor variable A on X is non-zero, invariance is attainable.

Summarizing, in these examples, *anchor regression* exhibits constant predictive performance even under arbitrarily strong shift interventions. In Section 8.3 we investigate the phenomenon of "invariance under interventions".

8.3 Data-driven invariance

In Section 8.2 we discussed three examples for which the distribution of $Y - X^\intercal b^{\to\infty}$ under $\mathbb{P}_v$ is invariant under certain shift interventions v. Here and in the following, we tacitly assume that the limit $b^{\to\infty} := \lim_{\gamma\to\infty} b^\gamma$ exists.

We want to investigate the conditions under which we have invariance. Define $I := \{b \in \mathbb{R}^d : \mathbb{E}_{\text{train}}[A \cdot (Y - X^\intercal b)] = 0\}$. Then we have the following theorem.

Theorem 6. *Assume that the Gram matrix $\mathbb{E}[AA^\intercal]$ is positive definite. Then,*

$$b \in I \qquad \Longleftrightarrow \qquad Y - X^\intercal b \text{ under } \mathbb{P}_v \text{ has the same distribution for all } v \in \operatorname{span}(\mathbf{M}).$$

Note that if the set I is non-empty, then $b^{\to\infty} \in I$. Hence *anchor regression* will have this invariance property for $\gamma \to \infty$ if and only if there exists a b that has this property.

This invariance can be interpreted as follows. Assume we have an anchor $A \in \{-1, 1\}$ that represents data collected from two environments. Data from environment $A = 1$ and environment $A = -1$ differ by a shift intervention of $2\mathbf{M}$ on (X, Y, H). The theorem above tells us that for all $b \in I$ the residual distribution (that means, the distribution of $Y - X^\intercal b$) is invariant under $\mathbb{P}_v$ with $v = \alpha\mathbf{M}, \alpha \in \mathbb{R}$. In this sense, *anchor regression* with $\gamma \to \infty$ is invariant with respect to the heterogeneities that are observed in the training distribution (to be more precise, we obtain invariance of the residuals with respect to linear combinations of inhomogeneities in the training distribution, cf. Theorem 6).

In the following discussion we will make an assumption that facilitates interpretation of the span of the shift matrix $\mathbf{M}$, i.e. of $\operatorname{span}(\mathbf{M})$. Define $T := \{k : \mathbf{M}_{k,\bullet} \not\equiv 0\}$ as the rows of $\mathbf{M}$ that are not identically zero. In the following we will refer to T as *children of* A. Let the Gram matrix of $(\mathbf{M}A)_T$ be positive definite. In the following, we will call this the *full-rank assumption*. Then $\operatorname{span}(\mathbf{M}) = \{v : v_{-T} \equiv 0\}$, the set that contains arbitrary interventions on the children of A. In particular, for all $b \in \mathbb{R}^d$,

$$b \in I \qquad \Longleftrightarrow \qquad Y - X^\intercal b \text{ under } \mathbb{P}_v \text{ has the same distribution for all } v \in \mathbb{R}^{d+1+r} \text{ with } v_{-T} \equiv 0.$$

Hence the set $I = \{b : \mathbb{E}_{\text{train}}[A \cdot (Y - X^\intercal b)] = 0\}$ is exactly the set of vectors b for which $Y - X^\intercal b$ is invariant under interventions on the children of A. This has consequences for the interpretation of $b^{\to\infty}$. If I is nonempty, i.e., if invariance is attainable, then loosely speaking

$$b^{\to\infty} = \operatorname*{argmin}_b \mathbb{E}_{\text{train}}[(Y - X^\intercal b)^2] \text{ s.t. the distribution of } Y - X^\intercal b \text{ under } \mathbb{P}_v \text{ has invariant distribution}$$
$$\text{under shift interventions on the children of } A.$$

Figure 7: In the example on the left, the full-rank assumption holds. In the example on the right, the full-rank assumption does not hold. The deterministic shifts in C^γ for $\gamma = 1$ are visualized in Figure 8.

Note that $\lim_{\gamma\to\infty} b^\gamma$ may exist, even in cases where $\mathbb{P}_v$ is not invariant under shift interventions $v \in C^\gamma$. Under the assumptions of Theorem 1, we have

$$b^\gamma = \arg\min_b \sup_{v\in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

Thus, if b^γ converges for $\gamma \to \infty$, $b^{\to\infty}$ corresponds to the prediction rule that results in the least-growing worst-case prediction loss for $v \in C^\gamma$, $\gamma \to \infty$.

In the next discussion we will give two examples to shed some light on the full-rank assumption.

8.4 Shape of C^γ

In the preceding section we saw examples where A has only one child leading to very simple forms of C^γ. In this section we will discuss two slightly more involved examples, with two covariates (X_1, X_2) and one hidden confounder H. The examples are depicted in Figure 7. Invariance in the sense of Theorem 6 is only achievable for the graph on the right: It can be shown that $I = \emptyset$ for the graph on the left.

In both examples, assume that $\mathbb{E}_{\text{train}}[AA^\intercal] = \text{Id}$. Then, in the example on the left, we have

$$\mathbf{M} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \\ 2 & 0 \end{pmatrix}, \text{ and hence } C^1 = \left\{ v \in \mathbb{R}^4 \text{ such that } v_2 = v_3 = 0 \text{ and } \left(v_1 - \frac{v_4}{2}\right)^2 + \frac{v_4^2}{4} \le 1 \right\}.$$

v_1 corresponds to interventions on X_1, whereas v_4 corresponds to interventions on H. The set C^1 is visualized in Figure 8 on the left-hand side. As $v_2 = v_3 = 0$ for all $v \in C^1$, only the dimensions v_1 and v_4 (interventions on X_1 and H) are shown. The full-rank assumption holds, as

$$\mathbf{M}_{T,\bullet} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix} \text{ has full row-rank.}$$

On the right-hand side the situation is different, as we only have one anchor. Here,

$$\mathbf{M} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 2 \end{pmatrix}, \text{ hence } C^1 = \left\{ v \in \mathbb{R}^3 \text{ such that } v_2 = v_3 = 0, 2v_1 = v_4 \text{ and } v_1^2 \le 1 \right\}.$$

Analogously as above, the deterministic shifts in C^1 are visualized in Figure 8 on the right-hand side. The ellipsoid is degenerate and the full-rank assumption is not fullfilled as

$$\mathbf{M}_{T,\bullet} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ does not have full row-rank.}$$

If $(\mathbf{M}A)_T$ is degenerate, then we observe shifts only in certain linear subspaces of $\mathbb{R}^{|T|}$ and *anchor regression* optimizes the MSE only under these restricted interventions. It seems desirable to include as many anchors as possible to optimize predictive performance under a wide range of interventions. However, this comes at a cost. Adding anchors that correspond to shifts that will not occur in the test data set can result in overly conservative predictive performance.

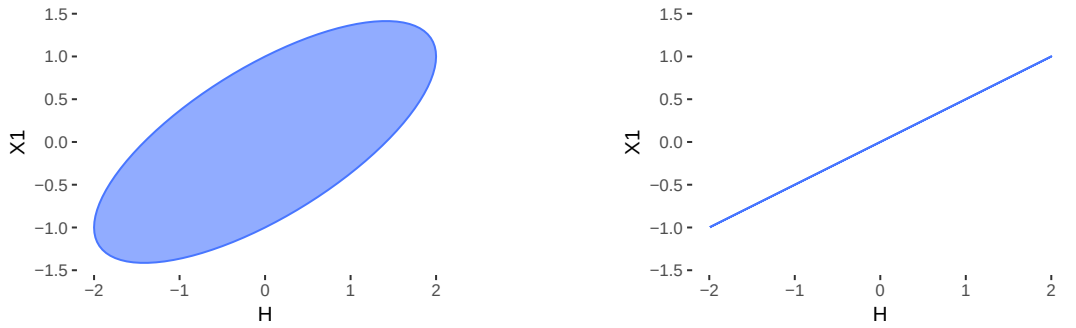


Figure 8: The blue areas correspond to the interventions in C^1 for the examples in Figure 7. On the left-hand side the full-rank assumption holds. Loosely speaking, for $\gamma \to \infty$ the ellipsoid grows larger and larger, eventually containing arbitrary shift interventions on X_1 and H. On the right-hand side, the full-rank assumption does not hold, hence C^γ for $\gamma \to \infty$ only contains interventions on X_1 and H that satisfy certain linear constraints.

8.5 Theorem 1 for random shifts

Theorem 7. *For any $b \in \mathbb{R}^d$ we have*

$$\mathbb{E}_{train}[((\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b))^2] + \gamma \mathbb{E}_{train}[(\mathrm{P}_A(Y - X^\intercal b))^2] = \sup_{\mathbb{P}_v \in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2], \tag{24}$$

where

$$C^\gamma := \{\textit{probability measures } \mathbb{P}_v : \textit{ the assumptions of Section 2.1 are satisfied, and } \mathbb{E}_v[vv^\intercal] \preceq \gamma \mathbf{M} \mathbb{E}_{train}[AA^\intercal]\mathbf{M}^\intercal\}.$$

8.6 Proof of Theorem 1 and Theorem 7

Proof. We will show Theorem 1. The proof of Theorem 7 proceeds analogously. Using the model assumptions of Section 2.1, under $\mathbb{P}_v$,

$$Y - X^\intercal b = ((\mathrm{Id} - \mathbf{B})^{-1}_{d+1,\bullet} - b^\intercal (\mathrm{Id} - \mathbf{B})^{-1}_{1:d,\bullet})(\varepsilon + v).$$

In the following, for brevity we write $w = ((\mathrm{Id} - \mathbf{B})^{-1}_{d+1,\bullet} - b^\intercal (\mathrm{Id} - \mathbf{B})^{-1}_{1:d,\bullet})^\intercal$. As $\mathbb{E}_v[\varepsilon] = 0$ and using that ε and v are uncorrelated under $\mathbb{P}_v$,

$$\mathbb{E}_v[(Y - X^\intercal b)^2] = \mathbb{E}_0[(Y - X^\intercal b)^2] + \mathbb{E}_v[(w^\intercal v)^2].$$

Taking the supremum over C^γ, using the definition of C^γ,

$$\begin{aligned}
\sup_{v \in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2] &= \mathbb{E}_0[(Y - X^\intercal b)^2] + \sup_{v \in C^\gamma} \mathbb{E}_v[(w^\intercal v)^2] \\
&= \mathbb{E}_0[(Y - X^\intercal b)^2] + \sup_{v \in C^\gamma} w^\intercal \mathbb{E}_v[vv^\intercal] w \\
&= \mathbb{E}_0[(Y - X^\intercal b)^2] + \gamma w^\intercal \mathbf{M} \mathbb{E}_{\mathrm{train}}[AA^\intercal]\mathbf{M}^\intercal w \\
&= \mathbb{E}_0[(Y - X^\intercal b)^2] + \gamma \mathbb{E}_{\mathrm{train}}[(w^\intercal \mathbf{M} A)^2]
\end{aligned} \tag{25}$$

By the model assumptions of Section 2.1, ε and A are independent and $\mathbb{E}_{\mathrm{train}}[\varepsilon] = 0$, which together with the definition of w implies that under $\mathbb{P}_{\mathrm{train}}$,

$$\begin{aligned}
\mathbb{E}_{\mathrm{train}}[Y - X^\intercal b | A] = \mathbb{E}_{\mathrm{train}}[w^\intercal(\varepsilon + \mathbf{M}A)|A] = w^\intercal \mathbf{M} A, \text{ and} \\
Y - X^\intercal b - \mathbb{E}_{\mathrm{train}}[Y - X^\intercal b | A] = w^\intercal(\varepsilon + \mathbf{M}A) - w^\intercal \mathbf{M} A = w^\intercal \varepsilon.
\end{aligned} \tag{26}$$

Note that by definition under $\mathbb{P}_0$, $Y - X^\intercal b$ has the same distribution as $w^\intercal \varepsilon$ under $\mathbb{P}_{\mathrm{train}}$. Hence, under $\mathbb{P}_{\mathrm{train}}$, $Y - X^\intercal b - \mathbb{E}_{\mathrm{train}}[Y - X^\intercal b | A]$ has the same distribution as $Y - X^\intercal b$ under $\mathbb{P}_0$. Thus, using the equations (26) in equation (25) yields

$$\sup_{v \in C^\gamma} \mathbb{E}_v[(Y - X^\intercal b)^2] = \mathbb{E}_{\mathrm{train}}[(Y - X^\intercal b - \mathbb{E}_{\mathrm{train}}[Y - X^\intercal b | A])^2] + \gamma \mathbb{E}_{\mathrm{train}}[(\mathbb{E}_{\mathrm{train}}[Y - X^\intercal b | A])^2],$$

which concludes the proof. □

8.7 Proof of Lemma 1

Proof. We can rewrite $\mathbb{E}[(Y - X^\intercal b)^2|A]$:

$$\mathbb{E}[(Y - X^\intercal b)^2|A] = \mathbb{E}[(Y - X^\intercal b - \mathbb{E}[Y - X^\intercal b|A])^2|A] + (\mathbb{E}[Y - X^\intercal b|A])^2$$

As (X, Y, A) follows a centered multivariate Gaussian distribution,

$$\mathbb{E}[Y - X^\intercal b|A] \sim \mathcal{N}(0, \mathbb{E}[(\mathbb{E}[Y - X^\intercal b|A])^2])$$
$$\text{and } \mathbb{E}[(Y - X^\intercal b - \mathbb{E}[Y - X^\intercal b|A])^2|A] = \mathbb{E}[(Y - X^\intercal b - \mathbb{E}[Y - X^\intercal b|A])^2].$$

Hence the α-th quantile of $\mathbb{E}[(Y - X^\intercal b)^2|A]$ is equal to

$$\mathbb{E}[(Y - X^\intercal b - \mathbb{E}[Y - X^\intercal b|A])^2] + \chi_1^2(\alpha)\mathbb{E}[(\mathbb{E}[Y - X^\intercal b|A])^2],$$

where $\chi_1^2(\alpha)$ denotes the α-th quantile of a χ^2-distributed random variable with one degree of freedom. □

8.8 Lemma 1 for discrete anchors

Lemma 3 (Version of Lemma 1 for discrete anchors)**.** *Assume that we have several training data sets $a \in \mathcal{A}$ and one test dataset. For each $a \in \mathcal{A}$, the data are drawn i.i.d. from $\mathbb{P}^a = \mathbb{P}[\bullet|A = a]$. On the test data set, the data are drawn i.i.d. from the distribution of $\mathbb{P}^{test}$. We write $\mathbb{E}^a[\bullet]$ to denote the expectation on data set $a \in \mathcal{A}$ and $\mathbb{E}^{test}$ to denote the expectation on the test data set. We assume that the data sets differ by a shift $\delta^a := \mathbb{E}^a[(X, Y)]$, i.e., that $(X, Y) - \mathbb{E}^a[(X, Y)]$ under $\mathbb{P}^a$ has the same distribution as $(X, Y) - \mathbb{E}^{a'}[(X, Y)]$ under $\mathbb{P}^{a'}$ for all $a \in \mathcal{A} \cup \{new\}$. We assume that the shift δ^a is constant on each data set, but random between the data sets, with distribution $\delta^a \sim \mathcal{N}(0, \Sigma)$ for some positive semi-definite Σ. Write $\mathbb{E}^{a,\delta}$ for the expectation both with respect to the randomness of $\mathbb{P}^a$ and the randomness of the shift δ^a. Due to the randomness of δ^{new}, the risk $\mathbb{E}^{new}[(Y - X^\intercal b)^2]$ is random and we write $Q(\alpha)$ for the quantiles of the risk on the new data set. Then,*

$$Q(\alpha) = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \mathbb{E}^{a,\delta}[((\mathrm{Id} - \mathbb{E}^a)[Y - X^\intercal b])^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \mathbb{E}^{a,\delta}[(\mathbb{E}^a[Y - X^\intercal b])^2], \tag{27}$$

where γ equals the α-th quantile of a χ^2-distributed random variable with one degree of freedom.

Note that with $\mathbb{E}_{\text{train}} \equiv \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \mathbb{E}^{a,\delta}$ and $\mathbb{E}^a \equiv \mathbb{E}_{\text{train}}[\bullet|A = a]$ the right-hand side coincides with the anchor risk if each level of A has equal probability (or under a re-weighting such that each level of A has equal probability).

Proof. First, note that using a bias-variance decomposition we can rewrite the risk on the new data set as

$$\mathbb{E}^{\text{new}}[(Y - X^\intercal b)^2] = \mathbb{E}^{\text{new}}[(Y - \delta_{p+1}^{\text{new}} - (X - \delta_{1:p}^{\text{new}})^\intercal b)^2] + (\delta_{p+1}^{\text{new}} - (\delta_{1:p}^{\text{new}})^\intercal b)^2$$

By assumption, the first part of this term is constant and the second part is a centered Gaussian random variable. Hence, the α-Quantile of this term is

$$Q(\alpha) = \mathbb{E}^{\text{new}}[(Y - \delta_{p+1}^{\text{new}} - (X - \delta_{1:p}^{\text{new}})^\intercal b)^2] + \gamma\mathbb{E}^{\text{new},\delta}[(\delta_{p+1}^{\text{new}} - (\delta_{1:p}^{\text{new}})^\intercal b)^2] \tag{28}$$

Now using that the distribution of $(X, Y) - \delta^{\text{new}}$ under $\mathbb{P}^{\text{new}}$ is the same as the distribution of $(X, Y) - \delta^{\text{new}}$ under $\mathbb{P}^a$, we obtain for all $a \in \mathcal{A}$

$$\mathbb{E}^{\text{new}}[(Y - \delta_{p+1}^{\text{new}} - (X - \delta_{1:p}^{\text{new}})^\intercal b)^2] = \mathbb{E}^{\text{a},\delta}[(Y - \delta_{p+1}^{\text{a}} - (X - \delta_{1:p}^{\text{a}})^\intercal b)^2]. \tag{29}$$

Using that the δ^a are i.i.d.,

$$\mathbb{E}^{\text{new},\delta}[(\delta_{p+1}^{\text{new}} - (\delta_{1:p}^{\text{new}})^\intercal b)^2] = \mathbb{E}^{\text{a},\delta}[(\delta_{p+1}^{\text{a}} - (\delta_{1:p}^{\text{a}})^\intercal b)^2]. \tag{30}$$

Using equation (29) and equation (30) in equation (28) completes the proof. □

8.9 Proof of Lemma 2

Proof. Due to linearity of the model in (8), $\mathbb{E}_{\text{train}}[Y|A]$ and $\mathbb{E}_{\text{train}}[X|A]$ are linear functions of A whose coefficients are given by the least squares principle. That is:

$$\mathbb{E}_{\text{train}}[Y|A] = A^\intercal \mathbb{E}_{\text{train}}[AA^\intercal]^{-1}\mathbb{E}_{\text{train}}[AY],$$
$$\mathbb{E}_{\text{train}}[X^\intercal|A] = A^\intercal \mathbb{E}_{\text{train}}[AA^\intercal]^{-1}\mathbb{E}_{\text{train}}[AX^\intercal].$$

Thus, $\mathbb{E}_{\text{train}}[Y - X^\intercal b|A] = 0$ if and only if

$$\text{Cov}_{\text{train}}(A, X)b = \text{Cov}_{\text{train}}(A, Y), \tag{31}$$

where we used that $\mathbb{E}_{\text{train}}[AA^T]$ is invertible and covariances replace expectations since X and Y are assumed to have mean zero. Equation (31) is a linear system of equations in the variables b: by the Rouché-Capelli theorem, it has a solution if and only if

$$\text{rank}(\text{Cov}_{\text{train}}(A, X)) = \text{rank}\left(\text{Cov}_{\text{train}}(A, X)|\text{Cov}_{\text{train}}(A, Y)\right),$$

which completes the proof. □

8.10 Proof of Theorem 2

Proof. Due to the projectability condition in (13) and Lemma 2 we know that $I \neq \emptyset$. The projectability condition also holds for X', Y', A' since one can verify that the rank condition only depends on $\mathbf{B}$ and $\mathbf{M}$. Thus, we also have that $I' \neq \emptyset$.

(i) Characterization of $b^{\to\infty}$.
We first consider the residual term

$$\eta_b = Y - X^\intercal b$$

for any $b \in I$. Analogously as in the proof of Theorem 3 consider

$$w_b = \left((\text{Id} - \mathbf{B})^{-1}_{d+1,\bullet} - b^\intercal(\text{Id} - \mathbf{B})^{-1}_{1:d,\bullet}\right)^\intercal.$$

We then have that $\eta_b = w_b^\intercal(\varepsilon + \mathbf{M}(\kappa A + \xi))$. Since $b \in I$, we have that $\mathbb{E}_{\text{train}}[\eta_b|A] = 0$ and therefore $\mathbb{E}_{\text{train}}[\eta_b A^\intercal] = \mathbb{E}_{\text{train}}[w_b^\intercal(\mathbf{M}\kappa AA^\intercal)] = 0$ (where we used in the first equality relation that $\mathbb{E}_{\text{train}}[\xi] = \mathbb{E}_{\text{train}}[\varepsilon] = 0$ and ξ, A, ε are jointly independent). Since $\mathbb{E}_{\text{train}}[AA^\intercal]$ is invertible we obtain

$$w_b^\intercal \mathbf{M} = 0 \ \forall\ b \in I. \tag{32}$$

Therefore $\eta_b = w_b^\intercal \varepsilon$ and thus we have:

$$b^{\to\infty} = \text{argmin}_{b\in I}\mathbb{E}_{\text{train}}[\eta_b^2] = w_b^\intercal \Sigma_\varepsilon w_b, \tag{33}$$

where $\Sigma_\varepsilon = \text{Cov}(\varepsilon)$.

(i) Characterization of $b'^{\to\infty}$.
One can derive exactly along the same lines as above the analogue of (32):

$$w_b^\intercal \mathbf{M} = 0 \ \forall b \in I'. \tag{34}$$

Because of (34) and using (17) we have that

$$\begin{aligned} b'^{\to\infty} &= \text{argmin}_{b\in I'}\mathbb{E}[(Y' - (X')^\intercal b)^2] \\ &= \text{argmin}_{b\in I'}\mathbb{E}[(\eta'_b)^2] = \text{argmin}_{b\in I'} L w_b^\intercal \Sigma_\varepsilon w_b. \end{aligned} \tag{35}$$

This is to be compared with (33).

It remains to show that

$$I = I'. \tag{36}$$

"$\subseteq$": take $b \in I$. Then, by (32), $w_b^\intercal \mathbf{M} = 0$. Therefore

$$\mathbb{E}_{\text{test}}[\eta'_b|A'] = \mathbb{E}[w_b^\intercal \varepsilon'|A'] = 0$$

where the last equality follows by independence of ε' and A' and $\mathbb{E}_{\text{test}}[\varepsilon'] = 0$. Thus, $b \in I'$.
"$\supseteq$": take $b \in I'$. Then, by (34), $w_b^\intercal \mathbf{M} = 0$. Therefore

$$\mathbb{E}_{\text{train}}[\eta_b | A] = \mathbb{E}_{\text{train}}[w_b^\intercal \varepsilon | A] = 0$$

where the last equality follows by independence of ε and A and $\mathbb{E}_{\text{train}}[\varepsilon] = 0$. Thus, $b \in I$.
These two relations prove (36).

By (33), (35) and (36), we complete the proof of the theorem. □

8.11 Proof of Proposition 1

Proof. Define $f(b) := \mathbb{E}_{\text{train}}[(\mathrm{P}_A(Y - X^\intercal b))^2]$ and $g(b) := \mathbb{E}_{\text{train}}[((\mathrm{Id} - \mathrm{P}_A)(Y - X^\intercal b))^2]$. By assumption, $\partial_b f(b^0) = \partial_b f(b^\infty) = \partial_b g(b^0) = \partial_b g(b^\infty) = 0$. The objective functional of anchor regression for a fixed value of $\gamma \geq 0$ is $g(b) + \gamma f(b)$. Hence also the derivative of the objective functional at b^0 is zero. As the objective functional $g(b) + \gamma f(b)$ is convex in b, $b = b^0$ is a minimizer of the objective functional. This completes the proof. □

8.12 Proof of Theorem 3

Proof. Define $\eta = Y - X^\intercal b^0$. As $b^0 = b^{\to\infty}$, using equation (14), $\mathbb{E}_{\text{train}}[\eta | A] = 0$. This implies that

$$\mathbb{E}_{\text{train}}[\eta \cdot A] = \mathbb{E}_{\text{train}}[\mathbb{E}_{\text{train}}[\eta | A] \cdot A] = 0.$$

Define $w = ((\mathrm{Id} - \mathbf{B})^{-1}_{d+1,\bullet} - (b^0)^\intercal (\mathrm{Id} - \mathbf{B})^{-1}_{1:d,\bullet})^\intercal$. By using the model assumptions, under $\mathbb{P}_{\text{train}}$,

$$\eta = w^\intercal (\varepsilon + \mathbf{M} A)$$

Using $\mathbb{E}_{\text{train}}[w^\intercal \mathbf{M} A A^t] = \mathbb{E}_{\text{train}}[\eta \cdot A] = 0$ and that $\mathbb{E}_{\text{train}}[AA^t]$ is invertible, we have $w^\intercal \mathbf{M} = 0$. Now, let v be a random variable uncorrelated of ε that takes values in $\mathrm{span}(\mathbf{M})$. As $w^\intercal \mathbf{M} = 0$, $w^\intercal v = 0$. Thus, under $\mathbb{P}_v$,

$$Y - X^\intercal b^0 = w^\intercal (\varepsilon + v) = w^\intercal \varepsilon$$

Hence $Y - X^\intercal b^0$ has the same distribution under $\mathbb{P}_v$ as under $\mathbb{P}_0$. Thus, for all b we have

$$\mathbb{E}_v[(Y - X^\intercal b)^2] \geq \mathbb{E}_0[(Y - X^\intercal b)^2] \geq \mathbb{E}_0[(Y - X^\intercal b^0)^2] = \mathbb{E}_v[(Y - X^\intercal b^0)^2].$$

In the first step we used that v is uncorrelated of ε and equation (9). In the second step we used the definition of b^0. In the third step we used that $Y - X^\intercal b^0$ has the same distribution under $\mathbb{P}_v$ as under $\mathbb{P}_0$. Thus,

$$b^0 \in \operatorname{argmin} \mathbb{E}_v[(Y - X^\intercal b)^2].$$

This completes the proof. □

8.13 Generalized version of Theorem 4

In the following we will relax the assumptions to allow for endogeneous anchors.

8.13.1 Relaxed anchor assumptions

Let the distribution of (X, Y, H, A) under $\mathbb{P}_{\text{train}}$ be a solution of the SEM

$$\begin{pmatrix} X \\ Y \\ H \\ A \end{pmatrix} = \mathbf{B} \cdot \begin{pmatrix} X \\ Y \\ H \\ A \end{pmatrix} + \varepsilon. \tag{37}$$

where $\mathbf{B} \in \mathbb{R}^{(d+1+r+q)\times(d+1+r+q)}$ is an unknown constant matrix and the covariates X, the anchors $A \in \mathbb{R}^q$, the hidden variables $H \in \mathbb{R}^r$, and the noise $\varepsilon \in \mathbb{R}^{d+1+r}$ are random vectors. We assume that under $\mathbb{P}_{\text{train}}$, X and Y are centered to mean zero, that ε and A have finite second moments and that the components of ε are independent of each other. To make the distribution of X, Y, H, A well-defined, in the following we assume that $\mathrm{Id} - \mathbf{B}$ is invertible. The model induces a directed graph G, with the edges given by the following construction: For every $\mathbf{M}_{k,l} \neq 0$, a directed edge is drawn from A_l to the k-th variable in the $(d + 1 + r + q)$-dimensional vector (X, Y, H, A). Analogously, for every $\mathbf{B}_{k,l} \neq 0$, a directed edge is drawn from the l-th variable in (X, Y, H, A) to the k-th variable in (X, Y, H, A).

Theorem 8 (Anchor stability implies causality)**.** *Let the assumptions of Section 8.13.1 hold with an acyclic graph G, and assume the projectability condition* (13)*. Furthermore, assume that for every disjoint sets of variables $V_1, V_2, V_3 \subset (X, Y, H, A)$, V_1 is d-separated of V_2 in G given V_3 if and only if the partial correlation part.cor$(V_1, V_2|V_3) = 0$. Furthermore assume that for each X_k there exists k' such that $A_{k'} \to X_k$. If $b^{\to\infty} = b^0$, then*

$$b^{\to\infty} = b^0 = \partial_x \mathbb{E}[Y|do(X = x)], \tag{38}$$

where the do-operator $\mathbb{E}[\bullet|do(X = x)]$ is defined as in Pearl [2009], Chapter 1. In addition, there is no confounder between X and Y, i.e., there is no H_k that is both an ancestor of some $X_{k'}$ and Y in G.

8.14 Proof of Theorem 4 and Theorem 8

Proof. The proof for both theorems proceeds analogously. In the following, the covariances and partial correlations are meant with respect to the measure $\mathbb{P}_{\text{train}}$. Define $\eta = Y - X^\intercal b^0$. As $b^0 = b^{\to\infty}$, using equation (14), $\mathbb{E}_{\text{train}}[\eta|A] = 0$. This implies that

$$\mathbb{E}_{\text{train}}[\eta \cdot A] = \mathbb{E}_{\text{train}}[\mathbb{E}_{\text{train}}[\eta|A] \cdot A] = 0.$$

As $\mathbb{E}_{\text{train}}[\eta \cdot A] = 0$ and η is centered, we have $\text{Cov}(Y - X^\intercal b^0, A) = 0$. Using Proposition 1, $b^1 = b^0$. Let us write b' for the linear regression coefficient of A on X. We have $0 = \text{Cov}(Y - X^\intercal b^1, A) = \text{Cov}(Y - X^\intercal b^1, A - X^\intercal b')$. Thus, by the definition of partial correlation, part.cor$(Y, A|X) = 0$. By assumption this implies that Y and A are d-separated given X in G.

We want to show that every backdoor path from X to Y is blocked given the empty set. If we can show this, by the Backdoor-Criterion [Pearl, 2009], due to linearity, b^1 is equal to the causal effect $\partial_x \mathbb{E}[Y|do(X = x)]$. As we showed that $b^1 = b^0$, this would imply the claim of the theorem. Hence it suffices to show that every backdoor path from X to Y is blocked given the empty set.

Step 1: First, we note that no descendant of Y can be in X. We will prove this by contradiction. Choose k such that X_k is a descendant of Y and maximal in the sense that no other $X_{k'}$, $k' \neq k$ is a descendant of Y and an ancestor of X_k. By construction, there exists a directed path $X_k \leftarrow \ldots \leftarrow Y$ such that the nodes on this path do not lie in X. The nodes on this path do also not lie in A as A is d-separated of Y given X. By assumption there exists a k' such that $A_{k'} \to X_k$. Hence there exists a path $A_{k'} \to X_k \leftarrow \ldots \leftarrow Y$ that is open given X. Hence A is not d-separated of Y given X, contradiction.

Step 2: Assume there exists a backdoor path from X to Y that is open given the empty set. As the path from X to Y is open given the empty set, it cannot contain a collider. Let this path starts at X_k.

We have shown that the path does not contain a collider, and by Step 1, X_k is not a descendant of Y. As the backdoor path is open given the empty set, it must be of the form $X_k \leftarrow ... \leftarrow Z \to \ldots \to Y$ and the nodes on the path do not lie in X. No node of the path lies in A as we showed that A is d-separated of Y given X. To sum it up, we can assume that no node on the path lies in A or X. However, we assumed that there exists k' such that $A_{k'} \to X_k$. This gives us a path $A_{k'} \to X_k \leftarrow \ldots \leftarrow Z \to \ldots \to Y$ from $A_{k'}$ to Y that is open given X. Contradiction! Hence, every backdoor path from X to Y is blocked given the empty set. By the Backdoor-Criterion [Pearl, 2009], due to linearity, b^1 is equal to the causal effect $\partial_x \mathbb{E}[Y|do(X = x)]$. As we showed that $b^1 = b^0$, the claim of the theorem follows. □

8.15 Proof of Theorem 5 and auxiliary results

Notation. Define the "residuals" $\mathbf{Z}^{(a)} := \mathbf{Y}^{(a)} - \mathbf{X}^{(a)} b^\gamma$ for all $a \in \mathcal{A}$. We write $\overline{\mathbf{X}}^{(a)}$ for the empirical mean of $\mathbf{X}^a$, i.e., $\overline{\mathbf{X}}^{(a)} := \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{X}^{(a)}_{i,\bullet}$. Analogously define $\overline{\mathbf{Y}}^{(a)} := \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{Y}^{(a)}_i$ and $\overline{\mathbf{Z}}^{(a)} := \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{Z}^{(a)}_i$. Additionally define the conditional means $\mu^{(a)}_{\mathbf{X}} := \mathbb{E}_{\text{train}}[X|A = a]$, $\mu^{(a)}_{\mathbf{Y}} := \mathbb{E}_{\text{train}}[Y|A = a]$ and $\mu^{(a)}_{\mathbf{Z}} := \mathbb{E}_{\text{train}}[Y - X^\intercal b^\gamma|A = a]$ for $a \in \mathcal{A}$.

8.15.1 Proof of Theorem 5

Proof. **Preliminaries.** We want to derive bounds for

$$R(\hat{b}) - \min_b R(b),$$

where

$$R(b) = \mathbb{E}_{\text{train}}[(Y - \mathbb{E}_{\text{train}}[Y|A] - (X - \mathbb{E}_{\text{train}}[X|A])^\intercal b)^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mathbb{E}_{\text{train}}[Y|A=a] - \mathbb{E}_{\text{train}}[X|A=a]^\intercal b)^2.$$

and

$$\hat{b} = \operatorname*{argmin}_b \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left(\mathbf{Y}_i^{(a)} - \overline{\mathbf{Y}}^{(a)} - (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})b\right)^2 + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\left(\overline{\mathbf{Y}}^{(a)} - \overline{\mathbf{X}}^{(a)}b\right)^2 + 2\lambda\|b\|_1.$$

Using the assumptions of Section 2.1, (Y, X) has the same distribution under $\mathbb{P}_0$ as $(Y - \mathbb{E}_{\text{train}}[Y|A], X - \mathbb{E}_{\text{train}}[X|A])$ under $\mathbb{P}_{\text{train}}$. Hence with $\mu_{\mathbf{X}}^{(a)} = \mathbb{E}_{\text{train}}[X|A=a]$ and $\mu_{\mathbf{Y}}^{(a)} = \mathbb{E}_{\text{train}}[Y|A=a]$ we can rewrite the risk as

$$R(b) = \mathbb{E}_0[(Y - X^\intercal b)^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mu_{\mathbf{Y}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal b)^2.$$

Step 1: rewriting the excess risk. By definition we have $b^\gamma = \operatorname{argmin}_b R(b)$. For all $b \in \mathbb{R}^d$, all $\lambda \geq 0$, and all $\gamma \geq 0$ we thus have the decomposition

$$\begin{aligned}\mathbb{E}_0[(Y - X^\intercal b)^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mu_{\mathbf{Y}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal b)^2 &= \mathbb{E}_0[(X(b-b^\gamma))^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}((\mu_{\mathbf{X}}^{(a)})^\intercal(b-b^\gamma))^2 \\ &\quad + \mathbb{E}_0[(Y - X^\intercal b^\gamma)^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\mu_{\mathbf{Y}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal b^\gamma)^2.\end{aligned}$$

Hence if we write $W(b) = \mathbb{E}_0[(X^\intercal(b-b^\gamma))^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}((\mu_{\mathbf{X}}^{(a)})^\intercal(b-b^\gamma))^2$, we can rewrite the excess risk as

$$R(\hat{b}) - \min_b R(b) = W(\hat{b}).$$

We want to relate this excess risk to the empirical excess risk. Define

$$\hat{W}(b) := \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})(b-b^\gamma)\right)^2 + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)}(b-b^\gamma))^2.$$

As $(X, Y, H)^\intercal = (\mathrm{Id} - \mathbf{B})^{-1}\varepsilon$ under $\mathbb{P}_0$ and ε follows a centered multivariate Gaussian distribution, under $\mathbb{P}_0$, (X, Y) follows a centered multivariate Gaussian distribution as well. Recall that in the proof of Theorem 1 we have shown that the distribution of (Y, X) under $\mathbb{P}_0$ is the same as $(Y - \mathbb{E}[Y|A], X - \mathbb{E}[X|A]) = (Y - \mu_{\mathbf{Y}}^A, X - \mu_{\mathbf{X}}^A)$ under $\mathbb{P}_{\text{train}}$. As A and ε are independent, the distribution of $(Y, X)|(A=a)$ under $\mathbb{P}_{\text{train}}$ is the same as the distribution of $(Y + \mu_{\mathbf{Y}}^{(a)}, X + \mu_X^{(a)})$ under $\mathbb{P}_0$. Using Lemma 4, we obtain that with probability exceeding $1 - 4\exp(-t)$,

$$W(\hat{b}) \leq \frac{C''}{|S^*|}\|\hat{b} - b^\gamma\|_1^2 + \hat{W}(\hat{b}), \tag{39}$$

where C'' is a constant that depends on c', $\max_k \operatorname{Var}(X_k^0)$, $\max_{a\in\mathcal{A}}\|\mu_{\mathbf{X}}^{(a)}\|_\infty$ and γ.

Step 2: bounds for empirical excess risk $\hat{W}(\hat{b})$ and $\|\hat{b} - b^\gamma\|_1$. It turns out that it is straightforward to derive finite-sample bounds for $\hat{W}(\hat{b})$ and $\|\hat{b} - b^\gamma\|_1$, leveraging existing finite-sample bounds for the Lasso. To this end, let us define

$$z^* := \left\|\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal\left(\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)}\right) + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)})^\intercal\overline{\mathbf{Z}}^{(a)}\right\|_\infty$$

From Lemma 5 it follows that with probability $1 - 6\exp(-t)$ we have $2\|z^*\|_\infty \leq \lambda$. Now we can use Lemma 6 to bound $\hat{W}(\hat{b})$ and $\|\hat{b} - b^\gamma\|_1$. Lemma 6 follows directly from Theorem 2.2 in van de Geer [2016], but the notation is different. Details can be found in Section 8.15.4. Use Lemma 6 with $b = b^\gamma$, $S = S^*$, $\lambda_\varepsilon = \|z^*\|_\infty$ and $\delta = 0.5$. This gives $\underline{\lambda} = \lambda - \|z^*\|_\infty \geq \frac{\lambda}{2}$, $\overline{\lambda} = 1.5\lambda + 0.5\|z^*\|_\infty \leq 2\lambda$ and $L \leq 8$ to obtain

$$\begin{aligned}\hat{W}(\hat{b}) &\leq \frac{4\lambda^2|S^*|}{\hat{\phi}^2(8, S)}, \\ \text{and } \|\hat{b} - b^\gamma\|_1 &\leq \frac{8\lambda|S^*|}{\hat{\phi}^2(8, S^*)}.\end{aligned}$$

Combining these two bounds with equation (39) yields the desired result. □

8.15.2 Lemma 4

Lemma 4. *Let X follow a centered multivariate Gaussian distribution under $\mathbb{P}_0$. Let $\mathbf{X}_{i,\bullet}^{(a)}$, $i = 1, \ldots, n_a$ be i.i.d. random variables that have the same distribution as $\mu_{\mathbf{X}}^{(a)} + X$ for some deterministic vectors $\mu_{\mathbf{X}}^{(a)} \in \mathbb{R}^d$ for $a \in \mathcal{A}$. Let $\sigma_{\max}^2 := \max_k \mathrm{Var}(X_k)$, $n_{\min} := \min_{a\in\mathcal{A}} n_a$, $\mu_{\max} := \max_{a\in\mathcal{A}} \|\mu_{\mathbf{X}}^{(a)}\|_\infty$ and define the empirical means $\overline{\mathbf{X}}^{(a)} := \frac{1}{n_a}\sum_{i=1}^{n_a} \mathbf{X}_{i,\bullet}^{(a)}$ for $a \in \mathcal{A}$. Let $t \geq 0$ such that*

$$|S^*|^2(t + \log(d) + \log(|\mathcal{A}|))/n_{\min} \leq c', \tag{40}$$

for some constant $c' > 0$. Then, with probability exceeding $1 - 4\exp(-t)$, for any vectors $b, b^\gamma \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}_0[(X^\intercal(b - b^\gamma))^2] + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}((\mu_{\mathbf{X}}^{(a)})^\intercal(b - b^\gamma))^2 \\
\leq &\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})(b - b^\gamma)\right)^2 + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)}(b - b^\gamma))^2 + \frac{C''}{|S^*|}\|b - b^\gamma\|_1^2,
\end{aligned}$$

where C'' is a constant that depends on c', $\sigma_{\max}$, $\mu_{\max}$ and γ.

Proof. We will derive bounds for

$$\left|\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})(b - b^\gamma)\right)^2 - \mathbb{E}_0[(X^\intercal(b - b^\gamma))^2]\right| \tag{41}$$

and

$$\left|\frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)}(b - b^\gamma))^2 - \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}((\mu_{\mathbf{X}}^{(a)})^\intercal(b - b^\gamma))^2\right| \tag{42}$$

separately. By elementary linear algebra,

$$\begin{aligned}
&\left|\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})(b - b^\gamma)\right)^2 - \mathbb{E}_0[(X^\intercal(b - b^\gamma))^2]\right| \\
\leq &\|b - b^\gamma\|_1^2\left\|\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)}) - \mathbb{E}_0[XX^\intercal]\right\|_\infty
\end{aligned}$$

Now, using $\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)}) = 0$ repeatedly,

$$\begin{aligned}
\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal(\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)}) &= \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal(\mathbf{X}_{i,\bullet}^{(a)} - (\overline{\mathbf{X}}^{(a)})) \\
&= \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal(\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) \\
&\quad - \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal(\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)
\end{aligned} \tag{43}$$

We treat these two terms separately. First, using a sub-Gamma tail bound [Boucheron et al., 2013, Chapter 2], with probability exceeding $1 - 2\exp(-t)$,

$$\begin{aligned}
&\max_{a\in\mathcal{A}}\left\|\frac{1}{n_a}\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal(\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) - \mathbb{E}_0[XX^\intercal]\right\|_\infty \\
\leq &\sigma_{\max}^2\left(\sqrt{\frac{4t + 4\log(d^2 \cdot |\mathcal{A}|)}{n_{\min}}} + \frac{4t + 4\log(d^2 \cdot |\mathcal{A}|)}{n_{\min}}\right).
\end{aligned}$$

Using a sub-Gaussian tail bound [Boucheron et al., 2013, Chapter 2], with probability exceeding $1 - 2\exp(-t)$,

$$\max_{a\in\mathcal{A}}\|\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal\|_\infty \leq \sqrt{2\frac{\sigma_{\max}^2}{n_{\min}}(t + \log(d \cdot |\mathcal{A}|))}. \tag{44}$$

On this event,

$$\left\| \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) \right\|_\infty \leq \max_{a \in \mathcal{A}} \left\| (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) \right\|_\infty$$
$$\leq 2 \frac{\sigma_{\max}^2}{n_{\min}} (t + \log(d \cdot |\mathcal{A}|))$$

Using these two bounds in equation (43), we obtain the following bound for equation (41):

$$\begin{aligned}
&\left| \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} \left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})(b - b^\gamma) \right)^2 - \mathbb{E}_0[(X^\intercal (b - b^\gamma))^2] \right| \\
&\leq \|b - b^\gamma\|_1^2 \left\| \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)}) - \mathbb{E}_0[XX^\intercal] \right\|_\infty \qquad (45) \\
&\leq \|b - b^\gamma\|_1^2 \left(\sigma_{\max}^2 \left(\sqrt{\frac{4t + 4\log(d^2 \cdot |\mathcal{A}|)}{n_{\min}}} + \frac{4t + 4\log(d^2 \cdot |\mathcal{A}|)}{n_{\min}} \right) + 2 \frac{\sigma_{\max}^2}{n_{\min}} (t + \log(d \cdot |\mathcal{A}|)) \right)
\end{aligned}$$

Let us now treat equation (42). Analogously as above,

$$\begin{aligned}
&\left| \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\overline{\mathbf{X}}^{(a)} (b - b^\gamma))^2 - \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} ((\mu_{\mathbf{X}}^{(a)})^\intercal (b - b^\gamma))^2 \right| \\
&\leq \gamma \|b - b^\gamma\|_1^2 \max_{a \in \mathcal{A}} \left\| (\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{X}}^{(a)} - \mu_{\mathbf{X}}^{(a)} (\mu_{\mathbf{X}}^{(a)})^\intercal \right\|_\infty . \qquad (46)
\end{aligned}$$

Again, we can use a decomposition

$$\begin{aligned}
(\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{X}}^{(a)} - \mu_{\mathbf{X}}^{(a)} (\mu_{\mathbf{X}}^{(a)})^\intercal &= (\overline{\mathbf{X}}^{(a)})^\intercal (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) + (\overline{\mathbf{X}}^{(a)})^\intercal (\mu_{\mathbf{X}}^{(a)})^\intercal - \mu_{\mathbf{X}}^{(a)} (\mu_{\mathbf{X}}^{(a)})^\intercal \\
&= (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) + \mu_{\mathbf{X}}^{(a)} (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) \\
&\quad + (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mu_{\mathbf{X}}^{(a)})^\intercal
\end{aligned}$$

Using this decomposition in equation (46),

$$\begin{aligned}
&\left| \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\overline{\mathbf{X}}^{(a)} (b - b^\gamma))^2 - \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} ((\mu_{\mathbf{X}}^{(a)})^\intercal (b - b^\gamma))^2 \right| \\
&\leq \gamma \|b - b^\gamma\|_1^2 \max_{a \in \mathcal{A}} \Big(\|(\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)\|_\infty + \|\mu_{\mathbf{X}}^{(a)} (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)\|_\infty \\
&\quad + \|(\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mu_{\mathbf{X}}^{(a)})^\intercal\|_\infty \Big)
\end{aligned}$$

Recall that $\mu_{\max} = \max_{a \in \mathcal{A}} \|\mu_{\mathbf{X}}^{(a)}\|_\infty$. Using the sub-Gaussian tail bound of equation (44) for all three terms in the preceding equation, we obtain the following bound:

$$\begin{aligned}
&\left| \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} (\overline{\mathbf{X}}^{(a)} (b - b^\gamma))^2 - \frac{\gamma}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} ((\mu_{\mathbf{X}}^{(a)})^\intercal (b - b^\gamma))^2 \right| \\
&\leq \gamma \|b - b^\gamma\|_1^2 3 \max \left(\mu_{\max} \sqrt{2 \frac{\sigma_{\max}^2}{n_{\min}} (t + \log(d \cdot |\mathcal{A}|))}, 2 \frac{\sigma_{\max}^2}{n_{\min}} (t + \log(d \cdot |\mathcal{A}|)) \right) \qquad (47)
\end{aligned}$$

Using equation (40) in equation (47) and equation (45) yields the desired result. □

8.15.3 Lemma 5

Lemma 5. *Let (X, Y) follow a centered multivariate Gaussian distribution under $\mathbb{P}_0$. Let $\mathcal{A}$ be a finite set and $(\mathbf{X}_{i,\bullet}^{(a)}, \mathbf{Y}_i^{(a)})$, $i = 1, \ldots, n_a$ i.i.d. observations that have the same distribution as $(\mu_{\mathbf{X}}^{(a)} + X, \mu_{\mathbf{Y}}^{(a)} + Y)$ for some deterministic quantities $\mu_{\mathbf{X}}^{(a)} \in \mathbb{R}^d$ and $\mu_{\mathbf{Y}}^{(a)} \in \mathbb{R}$ for $a \in \mathcal{A}$. Let $b^\gamma \in \mathbb{R}^d$ such that*

$$b^\gamma = \operatorname*{argmin}_b \mathbb{E}_0 \left[\sum_{a \in \mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{Y}_i^{(a)} - \mu_{\mathbf{Y}}^{(a)} - (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal) b)^2 \right] + \gamma \sum_{a \in \mathcal{A}} (\mu_{\mathbf{Y}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal b)^2. \qquad (48)$$

Define $n_{\min} := \min_{a\in\mathcal{A}} n_a$. *Let* $t \geq 0$ *such that*

$$\frac{t + \log(d \cdot |\mathcal{A}|)}{n_{min}} \leq c'$$

for some constant $c' > 0$. *Then, with probability exceeding* $1 - 6\exp(-t)$,

$$\begin{aligned} z^* &= \left\| \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal \left(\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)} \right) + \frac{\gamma}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} (\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{Z}}^{(a)} \right\|_\infty \\ &\leq \frac{C}{2} \sqrt{\frac{t + \log(d \cdot |\mathcal{A}|)}{n_{min}}}, \end{aligned} \tag{49}$$

where $\mathbf{Z}^{(a)} = \mathbf{Y}^{(a)} - \mathbf{X}^{(a)} b^\gamma$ *and* $\overline{\mathbf{Z}}^{(a)} = \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{Z}_i^{(a)}$. *The constant* C *depends on* $\mu_{\max}$, $\sigma_{\max}$, γ *and* c'. *Here,* $\sigma_{\max}$ *denotes the maximal standard deviation, i.e.,* $\sigma_{\max}^2 := \max(\max_k \operatorname{Var}(X_k), \operatorname{Var}(Y - X^\intercal b^\gamma))$ *and* $\mu_{\max}$ *denotes the maximal mean, i.e.,* $\mu_{\max} := \max(\max_{a\in\mathcal{A}} \|\mu_{\mathbf{X}}^{(a)}\|_\infty, |\mu_{\mathbf{Y}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal b^\gamma|)$.

Proof. Recall that $\mu_{\mathbf{Z}}^{(a)} = \mu_{\mathbf{Y}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal b^\gamma$ for $a \in \mathcal{A}$. By taking the derivative of the objective functional in equation (48) with respect to b,

$$\mathbb{E}_0 \left[\sum_{a\in\mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal \left(\mathbf{Z}_i^{(a)} - \mu_{\mathbf{Z}}^{(a)} \right) \right] + \gamma \sum_{a\in\mathcal{A}} \mu_{\mathbf{X}}^{(a)} \mu_{\mathbf{Z}}^{(a)} = 0.$$

Using this, we can decompose equation (49):

$$\begin{aligned} &\left\| \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal \left(\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)} \right) + \frac{\gamma}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} (\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{Z}}^{(a)} \right\|_\infty \\ &\leq \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \left\| \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal \left(\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)} \right) - \operatorname{Cov}(\mathbf{X}_{i,\bullet}^{(a)}, \mathbf{Z}_i^{(a)}) \right\|_\infty \\ &+ \frac{\gamma}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \left\| (\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{Z}}^{(a)} - \mu_{\mathbf{X}}^{(a)} \mu_{\mathbf{Z}}^{(a)} \right\|_\infty \end{aligned} \tag{50}$$

As $\sum_{i=1}^{n_a} (\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)}) = 0$ and $\sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)}) = 0$,

$$\begin{aligned} &\frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal (\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)}) \\ &= \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)}) \\ &= \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mathbf{Z}_i^{(a)} - \mu_{\mathbf{Z}}^{(a)}) + \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mu_{\mathbf{Z}}^{(a)} - \overline{\mathbf{Z}}^{(a)}) \\ &= \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mathbf{Z}_i^{(a)} - \mu_{\mathbf{Z}}^{(a)}) + (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mu_{\mathbf{Z}}^{(a)} - \overline{\mathbf{Z}}^{(a)}). \end{aligned}$$

Similarly,

$$\begin{aligned} &(\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{Z}}^{(a)} - \mu_{\mathbf{X}}^{(a)} \mu_{\mathbf{Z}}^{(a)} \\ = \;& (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\overline{\mathbf{Z}}^{(a)} - \mu_{\mathbf{Z}}^{(a)}) + (\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal \mu_{\mathbf{Z}}^{(a)} + \mu_{\mathbf{X}}^{(a)} (\overline{\mathbf{Z}}^{(a)} - \mu_{\mathbf{Z}}^{(a)}). \end{aligned}$$

Combining these decompositions with equation (50),

$$\begin{aligned} &\left\| \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})^\intercal \left(\mathbf{Z}_i^{(a)} - \overline{\mathbf{Z}}^{(a)} \right) + \frac{\gamma}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} (\overline{\mathbf{X}}^{(a)})^\intercal \overline{\mathbf{Z}}^{(a)} \right\|_\infty \\ &\leq \max_{a\in\mathcal{A}} \left\| \frac{1}{n_a} \sum_{i=1}^{n_a} (\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mathbf{Z}_i^{(a)} - \mu_{\mathbf{Z}}^{(a)}) - \operatorname{Cov}(\mathbf{X}_{i,\bullet}^{(a)}, \mathbf{Z}_i^{(a)}) \right\|_\infty \\ &+ (\gamma + 1) \max_{a\in\mathcal{A}} \|(\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal (\mu_{\mathbf{Z}}^{(a)} - \overline{\mathbf{Z}}^{(a)})\|_\infty \\ &+ \gamma \max_{a\in\mathcal{A}} \|(\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal \mu_{\mathbf{Z}}^{(a)}\|_\infty \\ &+ \gamma \max_{a\in\mathcal{A}} \|\mu_{\mathbf{X}}^{(a)} (\overline{\mathbf{Z}}^{(a)} - \mu_{\mathbf{Z}}^{(a)})\|_\infty \end{aligned} \tag{51}$$

Using a sub-Gaussian tail bound [Boucheron et al., 2013, Chapter 2], with probability exceeding $1 - 4\exp(-t)$ we have

$$\max_{a\in\mathcal{A}} \max(\|\overline{\mathbf{X}}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal\|_\infty, |\overline{\mathbf{Z}}^{(a)} - \mu_{\mathbf{Z}}^{(a)}|) \leq \sqrt{\frac{2\sigma_{\max}^2(\log(d\cdot|\mathcal{A}|)+t)}{n_{\min}}},$$

With a sub-Gamma tail bound [Boucheron et al., 2013, Chapter 2], with probability exceeding $1 - 2\exp(-t)$ we have that

$$\begin{aligned} &\max_{a\in\mathcal{A}} \left\| \frac{1}{n_a}\sum_{i=1}^{n_a}(\mathbf{X}_{i,\bullet}^{(a)} - (\mu_{\mathbf{X}}^{(a)})^\intercal)^\intercal(\mathbf{Z}_i^{(a)} - \mu_{\mathbf{Z}}^{(a)}) - \mathrm{Cov}_{\mathrm{train}}(\mathbf{X}_{i,\bullet}^{(a)}, \mathbf{Z}_i^{(a)}) \right\|_\infty \\ \leq \quad & \sigma_{\max}^2 \left(\frac{4t + 4\log(d\cdot|\mathcal{A}|)}{n_{\min}} + \sqrt{\frac{4t+4\log(d\cdot|\mathcal{A}|)}{n_{\min}}} \right). \end{aligned}$$

Recall that by assumption

$$\frac{t + \log(d\cdot|\mathcal{A}|)}{n_{\min}} \leq c'.$$

Using these bounds in equation (51), we obtain that with probability exceeding $1 - 6\exp(-t)$

$$\|z^*\|_\infty \leq \frac{C}{2}\sqrt{\frac{t+\log(d\cdot|\mathcal{A}|)}{n_{\min}}},$$

where C depends on $\sigma_{\max}$, $\mu_{\max}$, c' and γ. □

8.15.4 Lemma 6

The following result provides a bound on $\|\hat{b} - b^\gamma\|_1$ and $\hat{W}(\hat{b})$, with $\hat{W}(\bullet)$ and $\hat{b}$ defined as in Section 8.15.1. It follows directly from Theorem 2.2 in van de Geer [2016], but the notation is different.

Lemma 6. *Let λ_ε satisfy $\lambda_\varepsilon \geq \|z^*\|_\infty$. Let $0 \leq \delta < 1$ be arbitrary and define for $\lambda > \lambda_\varepsilon$ and all $S \subseteq \{1,\ldots,d\}$*

$$\underline{\lambda} := \lambda - \lambda_\varepsilon, \qquad \overline{\lambda} := \lambda + \lambda_\varepsilon + \delta\underline{\lambda}, \qquad L := \frac{\overline{\lambda}}{(1-\delta)\underline{\lambda}},$$

$$\hat{\phi}^2(L,S) := \min_{\|b_S\|_1=1, \|b_{-S}\|_1 \leq L} |S| \left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \frac{1}{n_a}\sum_{i=1}^{n_a}\left((\mathbf{X}_i^{(a)} - \overline{\mathbf{X}}^{(a)})b\right)^2 + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)}b)^2 \right).$$

Then for all $b \in \mathbb{R}^d$ and all sets S,

$$2\delta\underline{\lambda}\|\hat{b} - b\|_1 + \hat{W}(\hat{b}) \leq \hat{W}(b) + \frac{\overline{\lambda}^2|S|}{\hat{\phi}^2(L,S)} + 4\lambda\|b_{-S}\|_1.$$

Proof. From a mathematical perspective, the proof of this result is immediate. However, it requires a change of notation. Define $\tilde{n} := \sum_{a\in\mathcal{A}} n_a + |\mathcal{A}|$. With some abuse of notation we can define $\tilde{\mathbf{Y}} \in \mathbb{R}^{\tilde{n}}$ as the row-wise concatenation of $\sqrt{\frac{\tilde{n}}{|\mathcal{A}|n_a}}(\mathbf{Y}^{(a)} - \overline{\mathbf{Y}}^{(a)}) \in \mathbb{R}^{n_a}$, $a \in \mathcal{A}$ and $\sqrt{\frac{\tilde{n}\gamma}{|\mathcal{A}|}}\cdot\overline{\mathbf{Y}}^{(a)} \in \mathbb{R}$, $a \in \mathcal{A}$. Analogously define $\tilde{\mathbf{X}} \in \mathbb{R}^{\tilde{n}\times d}$ as the row-wise concatenation of $\sqrt{\frac{\tilde{n}}{|\mathcal{A}|n_a}}(\mathbf{X}^{(a)} - \overline{\mathbf{X}}^{(a)})$, $a \in \mathcal{A}$ and $\sqrt{\frac{\tilde{n}\gamma}{|\mathcal{A}|}}\cdot\overline{\mathbf{Y}}^{(a)}$, $a \in \mathcal{A}$. Recall that

$$\hat{W}(b) = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\frac{1}{n_a}\sum_{i=1}^{n_a}\left((\mathbf{X}_{i,\bullet}^{(a)} - \overline{\mathbf{X}}^{(a)})(b - b^\gamma)\right)^2 + \frac{\gamma}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(\overline{\mathbf{X}}^{(a)}(b-b^\gamma))^2. \tag{52}$$

By this definition, we can rewrite $\hat{W}(b)$ as

$$\hat{W}(b) = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{\mathbf{X}}_{i,\bullet}(b-b^\gamma))^2.$$

Furthermore, *anchor regression* $\hat{b}$ minimizes the functional

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{\mathbf{Y}}_i - \tilde{\mathbf{X}}_{i,\bullet}b)^2 + 2\lambda\|b\|_1.$$

We can also rewrite z^* as

$$z^* = \frac{1}{\tilde{n}}\left\|\tilde{\mathbf{X}}^\intercal(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}b^\gamma)\right\|_\infty \tag{53}$$

and define $\tilde{\varepsilon} := \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}b^\gamma$. Now let us cite the following Theorem.

Theorem 9 (Theorem 2.2 in van de Geer [2016])**.** *Let λ_ε satisfy $\lambda_\varepsilon \geq \frac{1}{\tilde{n}}\left\|\tilde{\mathbf{X}}^\intercal\tilde{\varepsilon}\right\|_\infty$. Let $0 \leq \delta < 1$ be arbitrary and define for $\lambda > \lambda_\varepsilon$ and all sets $S \subseteq \{1,\dots,d\}$*

$$\underline{\lambda} := \lambda - \lambda_\varepsilon, \qquad \overline{\lambda} := \lambda + \lambda_\varepsilon + \delta\underline{\lambda}, \qquad L := \frac{\overline{\lambda}}{(1-\delta)\underline{\lambda}},$$

$$\hat{\phi}^2(L,S) := \min_{\|b_S\|_1=1, \|b_{-S}\|_1 \leq L} |S|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{\mathbf{X}}_{i,\bullet}(b_S - b_{-S}))^2.$$

Then for all b and all sets S,

$$2\delta\underline{\lambda}\|\hat{b} - b\|_1 + \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{\mathbf{X}}_{i,\bullet}(\hat{b} - b^\gamma))^2 \leq \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}(\tilde{\mathbf{X}}_{i,\bullet}(b^\gamma - b))^2 + \frac{\overline{\lambda}^2|S|}{\hat{\phi}^2(L,S)} + 4\lambda\|b_{-S}\|_1.$$

Using the above-mentioned change of notation concludes the proof of Lemma 6.

□

8.16 Proof of Theorem 6

Proof. Define $\mathbf{G} := \mathbb{E}_{\text{train}}[AA^\intercal]$. Recall that

$$I = \{b : \mathbb{E}_{\text{train}}[A \cdot (Y - X^\intercal b)] = 0\}$$

and define

$$\begin{aligned} J := \{b :\ & \text{for all } v \text{ in the span of } \mathbf{M} \text{ we have that} \\ & Y - X^\intercal b \text{ has the same distribution under } \mathbb{P}_v \text{ as under } \mathbb{P}_{\text{train}}\}. \end{aligned}$$

We will show $I = J$. For simplicity, in the following we will write $w_b := ((\mathrm{Id} - \mathbf{B})^{-1}_{d+1,\bullet} - b^\intercal(\mathrm{Id} - \mathbf{B})^{-1}_{1:d,\bullet})^\intercal$. Using the model assumptions of Section 2.1,

$$\begin{aligned} \mathbb{E}_{\text{train}}[A \cdot (Y - X^\intercal b)] &= \mathbb{E}_{\text{train}}\left[A \cdot (w_b^\intercal(\varepsilon + \mathbf{M}A))\right] \\ &= \mathbb{E}_{\text{train}}\left[A \cdot (w_b^\intercal\mathbf{M}A)\right]. \end{aligned}$$

As $\mathbb{E}_{\text{train}}[AA^\intercal] = \mathbf{G}$, it can be rewritten as $A = \mathbf{G}^{1/2}Z$ with $\mathbb{E}_{\text{train}}[ZZ^\intercal] = \mathrm{Id}$. Hence,

$$\begin{aligned} \mathbb{E}_{\text{train}}[A \cdot (Y - X^\intercal b)] &= \mathbb{E}_{\text{train}}\left[(\mathbf{G}^{1/2}Z) \cdot \left(w_b^\intercal\mathbf{M}\mathbf{G}^{1/2}Z\right)\right] \\ &= \mathbf{G}^{1/2}\mathbb{E}_{\text{train}}\left[Z \cdot \left(w_b^\intercal\mathbf{M}\mathbf{G}^{1/2}Z\right)\right] \\ &= \mathbf{G}^{1/2}\left(w_b^\intercal\mathbf{M}\mathbf{G}^{1/2}\right)^\intercal \end{aligned}$$

As $\mathbf{G}$ is assumed to be invertible, $\mathbb{E}_{\text{train}}[A \cdot (Y - X^\intercal b)] = 0$ if and only if $w_b^\intercal\mathbf{M} = 0$. This implies

$$I = \{b : w_b^\intercal\mathbf{M} = 0\}. \tag{54}$$

Using the model assumptions of Section 2.1, under $\mathbb{P}_v$,

$$\begin{aligned} Y - X^\intercal b &= w_b^\intercal(\varepsilon + v), \\ \text{and under } \mathbb{P}_{\text{train}} \text{ we have } Y - X^\intercal b &= w_b^\intercal(\varepsilon + \mathbf{M}A). \end{aligned}$$

The distributions of these random variables are equal for all $v \in \mathrm{span}(\mathbf{M})$ if and only if $w_b^\intercal\mathbf{M} = 0$. Hence,

$$J = \{b : w_b^\intercal\mathbf{M} = 0\}.$$

Using equation (54) concludes the proof. □

8.17 Figures for evaluating replicabilty

We show here additional results for replicability of variable selection in the GTEx data.

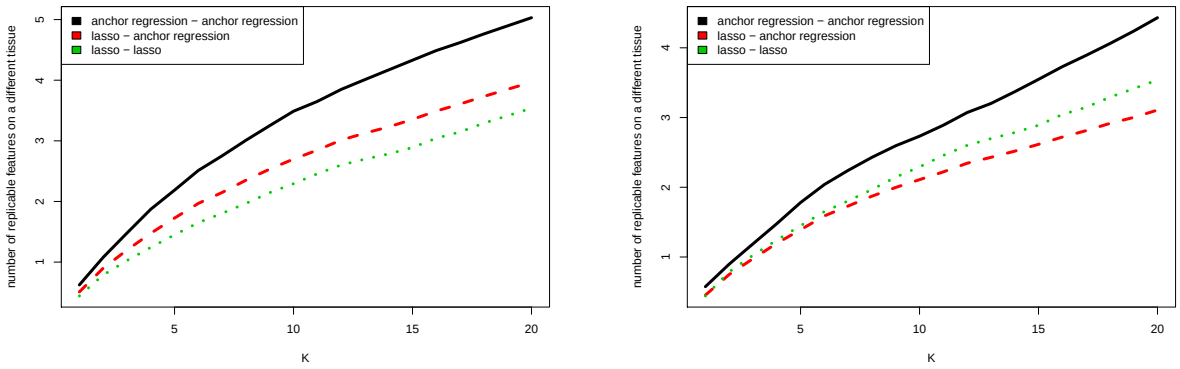


Figure 9: Replicability of variable selection on GTEx data. Same caption as in Figure 4, but now with $a_{y,k,t} := \min_{\gamma \in [0,.25]} |\hat{b}_k^{\gamma,\lambda}|$ on the left, and with $a_{y,k,t} := \min_{\gamma \in [0,16]} |\hat{b}_k^{\gamma,\lambda}|$ on the right.

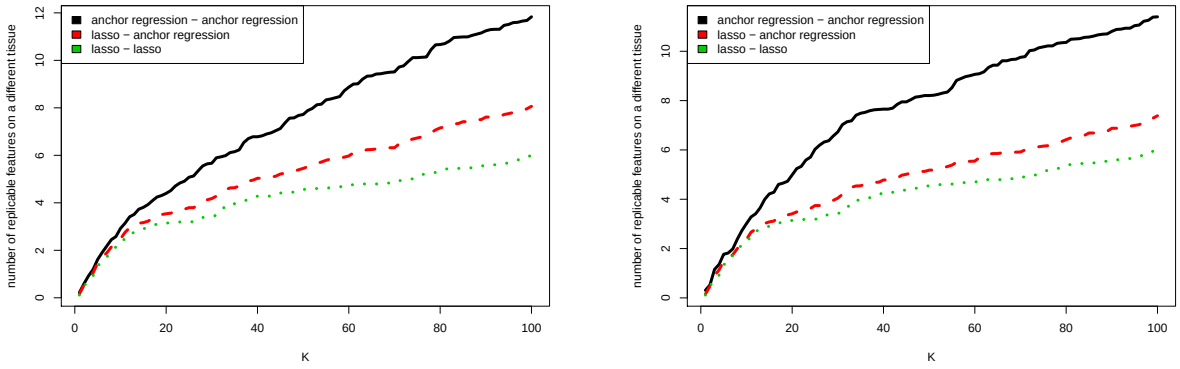


Figure 10: Replicability of variable selection on GTEx data. Same caption as in Figure 4, but with $a_{y,k,t} := \min_{\gamma \in [0,.25]} |\hat{b}_k^{\gamma,\lambda}|$ on the left, and with $a_{y,k,t} := \min_{\gamma \in [0,16]} |\hat{b}_k^{\gamma,\lambda}|$ on the right. Furthermore, the variable ranking is done over the 200 choices of the target variable y and averaging the results, instead of a fixed target y.

8.18 Figures for the bike sharing application

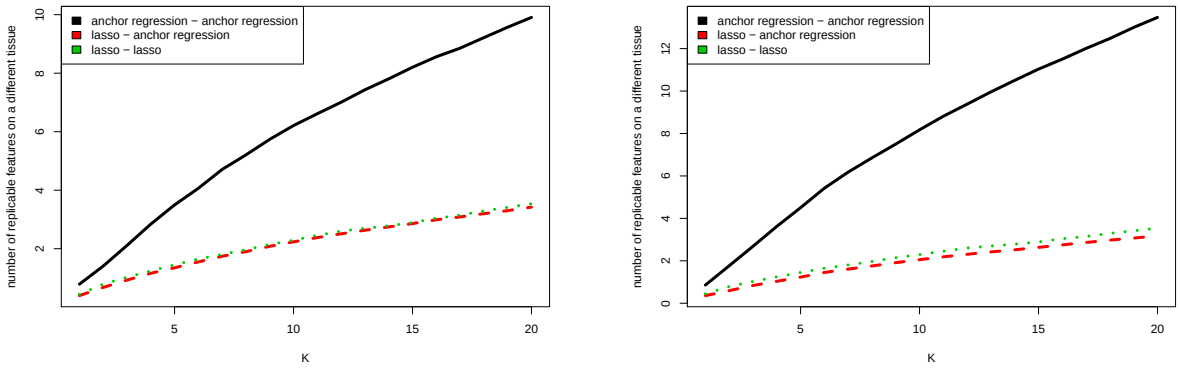


Figure 11: Replicability of variable selection on GTEx data. Same caption as in Figure 4, but with $a_{y,k,t} = |\hat{b}_k^{\gamma,\lambda}|$ for $\gamma = 8$ (left) and $\gamma = 16$ (right). While the coefficients show high replicability it depends on the interpretation of the anchor whether the coefficients are scientifically meaningful quantities. This is further discussed at the end of Section 3.1.

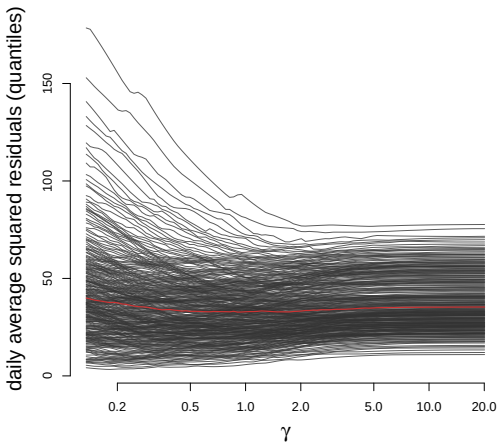


Figure 12: The plot is computed similarly as in Figure 5, but without removing the effect of working day, weekday and holiday in a pre-processing step. The two plots are very similar, i.e. practically it makes little difference whether the effect of working day, weekday and holiday are removed in a pre-processing step or not.

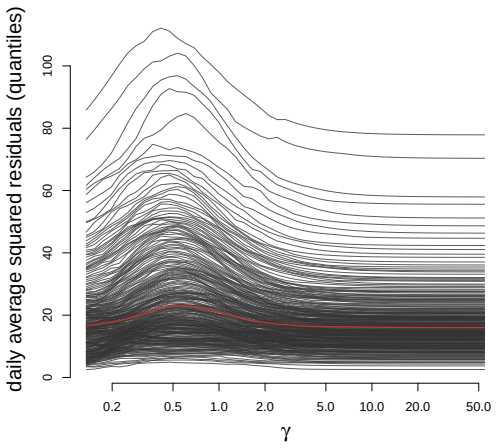


Figure 13: The plot is computed similarly as in Figure 5, but with the modified anchor procedure which is described at the end of Section 5.2. For large quantiles of the conditional loss, $\gamma \gg 1$ outperforms $\gamma < 1$, but the relationship is not monotonous and the performance of $\gamma \approx 0$ and $\gamma = 50$ are close.

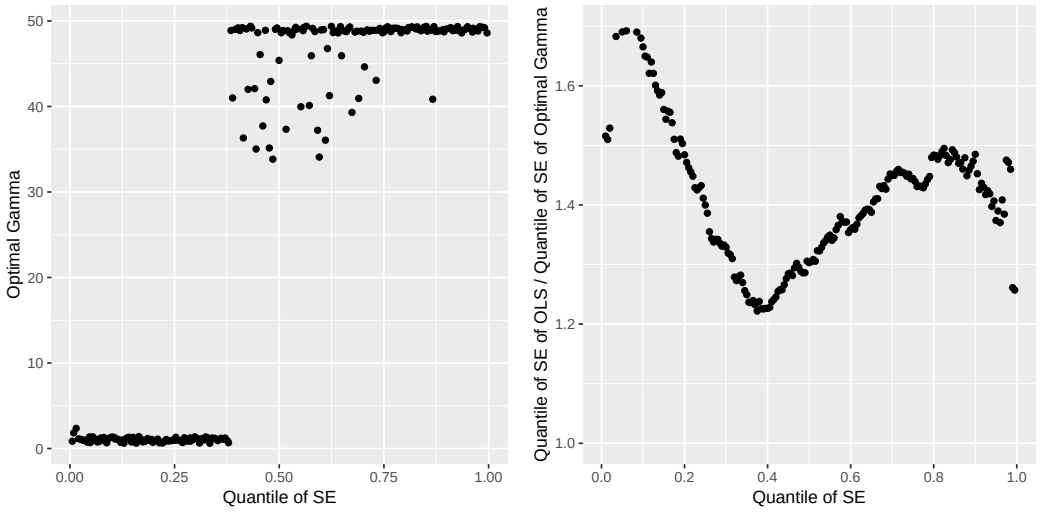


Figure 14: The plots were computed similarly as in Figure 6, but with the modified anchor procedure which is described at the end of Section 5.2. For small quantiles, $\gamma \approx 0$ is optimal, while for large quantiles $\gamma \approx 50$ is optimal. However, as can be seen in Figure 13, the performance of $\gamma = 0$ and $\gamma \approx 50$ are close. The anchor regression procedure performs better than ordinary least-squares for all considered quantiles.