Assessing the Overall and Partial Causal Well-Specification of Nonlinear Additive Noise Models

Christoph Schultheiss   CHRISTOPH.SCHULTHEISS@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
Zürich, 8092, CH

Peter Bühlmann   PETER.BUEHLMANN@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
Zürich, 8092, CH

Editor: Jin Tian

Abstract

We propose a method to detect model misspecifications in nonlinear causal additive and potentially heteroscedastic noise models. We aim to identify predictor variables for which we can infer the causal effect even in cases of such misspecification. We develop a general framework based on knowledge of the multivariate observational data distribution. We then propose an algorithm for finite sample data, discuss its asymptotic properties, and illustrate its performance on simulated and real data.

Keywords: Structural causal models, hidden variables, model misspecification, conditional independence, multi sample splitting

1. Introduction

Nonlinear additive noise models and their heteroscedastic extensions are a popular modelling framework for causal discovery and inference. They allow to infer the true causal connections and effects from the multivariate distribution when the nonparametric model is correct; see, e.g., Hoyer et al. (2008); Peters et al. (2014) or, for heteroscedastic models, Strobl and Lasko (2023); Immer et al. (2023). However, the conclusions can be misleading if the additive noise model is misspecified, especially in the presence of hidden confounding variables. In this paper, we define the term "causal well-specification" of additive noise models, discuss its relevance, and finally present a corresponding estimation technique for observational data.

The concept of well-specification for regression functionals in parametric regression was introduced by Buja et al. (2019). A regression functional is well-specified for a conditional target distribution if it only depends on the conditional distribution but is invariant to shifts in the predictors' distribution. This relates to the work by Peters et al. (2016) and, hence, gives the notion of well-specification a causal interpretation. Buja et al. (2019) suggest a set of reweighting diagnostics to assess well-specification of regression functions. For the linear model, an explicit test with asymptotic level as well as precise per-covariate interpretation for certain models is presented by Schultheiss et al. (2023).

If there is no functional assumption for the additive noise model, one must rely on flexible nonparametric regression techniques that approximate the conditional mean. Considering well-specification of the conditional mean is of little use. It is by definition a property of the conditional distribution only. Hence, it is, upon existence, well-specified for arbitrary data generating mechanisms.

Thus, different concepts are needed to infer whether the estimated effects in an assumed additive noise model are causal. One of our contributions is the definition of causal well-specification and presenting its interpretation. Apart from global causal well-specification, we also define a local, i.e., per predictor version that is to be considered when the overall model does not satisfy the desired properties. This local viewpoint is of particular interest in the presence of hidden common causes, hidden mediating variables, or misspecified functional form, e.g., the effect of the unmeasured independent error cannot be separated as an individual addend. We propose a methodology to assess causal well-specification from observational data by relying on and exploiting conditional independence. Based on this, we derive an algorithm for finite sample data and prove its consistency. From a practical viewpoint, our estimated set of well-specified predictors (i.e., covariables) can be viewed as the one where the data is compatible (i.e., does not falsify) with the corresponding local structure of the model and its (partial) causal interpretation.

Almost no work exists on local goodness of fit or well-specification of nonlinear causal models, where local well-specification has a causal interpretation. The latter is the main goal of the present paper. Our method works in arbitrary structural causal models, i.e., if there is no well-specification, the interpretation becomes conservative but not wrong.

Causal structure learning with hidden variables in greatest generality is treated within the framework of Fast Causal Inference (FCI, Spirtes, 2001). More specific and weakly related to our work is the approach by Maeda and Shimizu (2021) which discusses hidden variables in causal additive models (CAM, Bühlmann et al., 2014): unlike our current work, Maeda and Shimizu (2021) rely on the correctness of the causal additive model assumption. They present a causal graph detection algorithm based on unconditional independence tests. We do not provide a graph search technique but consider verification or falsification of assumed causal structures instead. We allow for as much flexibility in the model as possible while the CAM restricts the causal effect to sums of univariate functions. With the method in Maeda and Shimizu (2021) which is based on unconditional independence tests, certain edges remain undirected. By considering conditional independence, additional edges could be directed - at least with a conditional independence oracle which is at the basis of our approach. After introducing our theory, we present an example in Section 2.4 which illustrates some gains over the approach by Maeda and Shimizu (2021) by exploiting conditional independence.

2. Causal Well-Specification in Population

We consider first the population case in which we know the joint distribution of the observed random variables, e.g. conditional expectations and conditional independence between random variables can be perfectly assessed. This section is a stand-alone and can be used in connection with other estimation algorithms than the ones presented in Sections 3 and 6.1.

In Section 2.1, we introduce the causal model, our notation, and the most important background concepts from the causality literature. Section 2.2 provides a "roadmap" of our methodology. We describe on a high level which assumptions lead to a causal interpretation of the additive noise model, and how we can assess these using conditional independence statements. The mathematical details around these concepts follow in Sections 2.3 and 2.4.

2.1 Structural Causal Model

We summarize the concepts from the causality literature that are fundamental to our work. Let $\mathbf{Z} = (Z_1, \ldots, Z_q)^\top \in \mathbb{R}^q$ be a random vector whose entries Z_j follow a structural causal model (SCM), say, $\mathfrak{C}$,

$$\mathfrak{C}: \quad Z_j \leftarrow f_j\big(\mathbf{Z}_{\mathrm{PA}(j)}, \xi_j\big) \ \forall j \in \{1, \ldots, q\}. \tag{1}$$

We write $\leftarrow$ to emphasize that the equality is induced by a causal effect. ξ_j is some noise that is jointly independent over j. The set $\mathrm{PA}(j)$ denotes the parents of j, i.e., covariates Z_k with $k \in \mathrm{PA}(j)$ have a direct causal effect on Z_j. Conversely, j is a child of k. The SCM is represented by a directed graph that has an edge from $Z_k \to Z_j$ if and only if $k \in \mathrm{PA}(j)$. We assume that this results in a directed acyclic graph (DAG).

If the DAG contains any directed path from k to j, which may include several edges, we call j a descendant of k, $j \in \mathrm{DE}(k)$, and k an ancestor of j. On a path, $Z_k \to Z_l \to Z_j$, we call l a mediator. In a structure $Z_k \leftarrow Z_l \to Z_j$, l is a confounder.

We use the concept of d-separation (Geiger et al., 1990, Section 3). Two sets of variables $\mathbf{Z}_A$ and $\mathbf{Z}_B$ are d-separated by $\mathbf{Z}_C$ if it blocks all paths from $\mathbf{Z}_A$ to $\mathbf{Z}_B$. There are two ways to block a path:

- $\exists j \in C$ such that $Z_k \to Z_j \to Z_l$, $Z_k \leftarrow Z_j \leftarrow Z_l$, or $Z_k \leftarrow Z_j \to Z_l$ is on the path.

- $\exists j \notin C$ such that $Z_k \to Z_j \leftarrow Z_l$ is on the path, and $(\mathrm{DE}(j) \cap C) = \emptyset$.

The joint independence of the ξ_j in (1), implies that d-separated sets of variables are independent conditioned on the separating set (Pearl, 2009, Theorem 1.4.1). This is called the global Markov property. It applies for unconditional independence with $C = \emptyset$ as well. The distribution is called faithful to its DAG if all independences are implied by such a d-separation (Spirtes et al., 2000, Chapter 2.3.3). Violations of faithfulness can intuitively be described as cancellations of effects such that dependencies that one would assume to exist from the graph alone vanish.

We are interested in the situation where one variable with index in $\{1, \ldots, q\}$ is the target, some of the variables are observed (potential) predictors and the rest are unobserved or ignored (potential) predictors. Let Y, M (**m**easured) and N (**n**ot measured) be a partition of $\{1, \ldots, q\}$ that represent these subsets, and define the corresponding random variables and vectors

$$Y := Z_Y, \quad \mathbf{X} := \mathbf{Z}_M \in \mathbb{R}^p, \quad \mathbf{H} := \mathbf{Z}_N,$$
$$\mathbf{X}_{\mathrm{PA}(Y)} := \mathbf{Z}_{M \cap \mathrm{PA}(Y)} \quad \text{and} \quad \mathbf{H}_{\mathrm{PA}(Y)} := \mathbf{Z}_{N \cap \mathrm{PA}(Y)}.$$

In words, Y is the target, $\mathbf{X}$ are observed covariates, $\mathbf{H}$ are latent variables, $\mathbf{X}_{\mathrm{PA}(Y)}$ is the subset of Y's parents that we observe, and $\mathbf{H}_{\mathrm{PA}(Y)}$ is the subset that we do not observe.

With a slight abuse of notation, Y can represent the target random variable or the index in $\{1, \ldots, q\}$ that corresponds to the target. Note that for notational simplicity, we can absorb ξ_Y to be an additional variable in $\mathbf{H}_{\text{PA}(Y)}$. Therefore, $\mathbf{H}_{\text{PA}(Y)}$ always has dimensionality of at least one assuming Y is not deterministic in $\mathbf{X}$. In our SCM (1), we then have

$$Y \leftarrow f_Y\big(\mathbf{X}_{\text{PA}(Y)}, \mathbf{H}_{\text{PA}(Y)}\big).$$

For a realization $\mathbf{z}$ of $\mathbf{Z}$, we use the same naming convention, e.g., the realization of $\mathbf{X}$ is then $\mathbf{x}$.

We define the term Markov blanket. Consider $\mathbf{H}_{\text{PA}(Y)}$ as an exemplary target, analogous definitions for other targets exist. We call a set S a Markov blanket of these hidden parents if

$$\mathbf{H}_{\text{PA}(Y)} \perp \mathbf{X}_{-S} | \mathbf{X}_S,$$

where $\mathbf{X}_{-S}$ denotes all observed variables that are not in S. Importantly, we always mean these blankets to be found within only the observed covariates $\mathbf{X}$. Markov blankets are also known as sufficient sets. We define minimal Markov blankets as Markov boundaries, i.e., a set S such that

$$\mathbf{H}_{\text{PA}(Y)} \perp \mathbf{X}_{-S} | \mathbf{X}_S, \text{ but } \forall S' \subset S : \mathbf{H}_{\text{PA}(Y)} \not\perp \mathbf{X}_{-S'} | \mathbf{X}_{S'}.$$

As the Markov boundary is defined within only the observed covariates $\mathbf{X}$, in a structure $H_1 \rightarrow H_2 \rightarrow X_1$, X_1 would still count as part of the boundary of H_1 since it is the nearest measured descendant. We discuss the uniqueness of the Markov boundary in Section 2.4. It could also be all of $\mathbf{X}$ or empty.

We use causal do-notation to denote interventions. Conditioning on, e.g., $\text{do}(Z_j \leftarrow z_j)$ means that we assume a variation of the SCM $\mathfrak{C}$ (1) where the structural assignment for Z_j is not a function in its parents and the noise term but set to a fixed value. The remaining structural equations remain the same. Similarly, $\text{do}(\mathbf{Z}_S \leftarrow \mathbf{z}_S)$ means that a whole set of variables is intervened to have a fixed value.

We also apply the related concept of counterfactuals: what would happen to an observed data point if some of the covariates are set to hard values while the remaining structural assignments and unobserved noise terms remain unaffected? We use the notation from Chapter 6.4 in Peters et al. (2017), i.e., $P_Y^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\text{do}(\mathbf{X}\leftarrow\mathbf{x}')}$ denotes the counterfactual distribution of Y in the SCM $\mathfrak{C}$ where $\mathbf{Z} = \mathbf{z}$ is observed and the counterfactual intervention is $\mathbf{X} \leftarrow \mathbf{x}'$.

2.2 Roadmap of Our Methodology

We describe here on a high level the idea of causal well-specification of the general additive noise model as defined below. The interplay between the different assumptions, their causal implications, and how we aim to test for it is then visualized in Figure 1. Detailed mathematical definitions, assumptions, and results are given in the subsequent sections.

The additive noise model (ANM) has the following structure

$$Y \leftarrow f_{\mathbf{X}Y}\big(\mathbf{X}_{\text{PA}(Y)}\big) + f_{\mathbf{H}Y}\big(\mathbf{H}_{\text{PA}(Y)}\big), \quad \text{where} \quad \mathbf{X} \perp \mathbf{H}_{\text{PA}(Y)}.$$

4

It implies a testable proxy

$$H_0: \quad \mathcal{E} \perp \mathbf{X}, \quad \text{where} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}].$$

Independence from the hidden causes means that for the outcome's distribution it makes no difference if we observe $\mathbf{X} = \mathbf{x}$ or enforce it by intervention. With the additivity, we can even understand how the outcome reacts to a counterfactual change in $\mathbf{X}$ while keeping the unobserved ξ_k in (1) $\forall k \in N$ fixed. Hence, we can understand how the system reacts to change purely from the observational distribution.

The ANM assumption or the null-hypothesis H_0 can be violated in presence of dependence between $\mathbf{X}$ and hidden causal parents of Y, or misspecified functional form, meaning interactions between $\mathbf{X}_{\mathrm{PA}(Y)}$ and the hidden parents of Y in the structural equation for Y, or both. However, it can be that some observed covariates do not interact with the hidden causes, (A2) later, and are conditionally independent of these given the remaining covariates, (A1) later. We then say the ANM is causally well-specified for these covariates, say, $\mathbf{X}_U$. This implies another testable proxy

$$H_{0,U}: \quad \mathcal{E} \perp \mathbf{X}_U | \mathbf{X}_{-U}, \quad \text{with} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}].$$

As for all observed variables, M, we do not restrict U to be among the parents of Y since we typically do not have knowledge of these. Our causal interpretation of well-specification remains valid even for $j \in U \setminus \mathrm{PA}(Y)$: we can correctly characterize the absence of effects. Conditional independence from the unobserved causes means that for the outcome's distribution it makes no difference if we observe $\mathbf{X}_U = \mathbf{x}_U$ or enforce it by intervention at fixed levels of $\mathbf{X}_{-U}$. With the additivity, we can even understand how the outcome reacts to a counterfactual change in $\mathbf{X}_U$ while keeping $\mathbf{X}_{-U}$ and the unobserved ξ_k in (1) $\forall k \in N$ fixed. Hence, we can understand how the system reacts to changes in some covariates purely from the observational distribution.

By definition, such $\{1, \ldots, p\} \setminus U$ is a Markov blanket of $\mathcal{E}$. The more variables we can put into U the more explicative our model becomes. Hence, we aim to find a Markov boundary. But, having any blanket is enough to avoid false causal claims. We summarize in Figure 1.

Importantly, our results do not assume any model apart from the SCM in (1). We require independence and additivity to identify causal implications of the ANM. But, we can correctly falsify models if these assumptions do not hold.

Such causal well-specification can also be of use if one is mainly interested in purely predictive tasks and aims for out-of-distribution generalization where the new (test) data distribution is different ("shifted") from the training distribution; see, e.g., Rojas-Carulla et al. (2018) or the survey by Wang et al. (2023). Consider the task of domain adaptation with only few data in a target domain but many observations in a different training domain. If the distribution of the ξ_j in the target environment is shifted, also the best predictive function, $\mathbb{E}[Y|\mathbf{X}]$ might be different such that the large training set is not suitable for learning the predictive function. But, assuming invariant causal assignments in our SCM (1), the addend of the conditional mean induced by the causally well-specified covariates remains invariant across such shifted domains. Hence, this invariant part of the function could be estimated using the large multi-source data set from different domains.
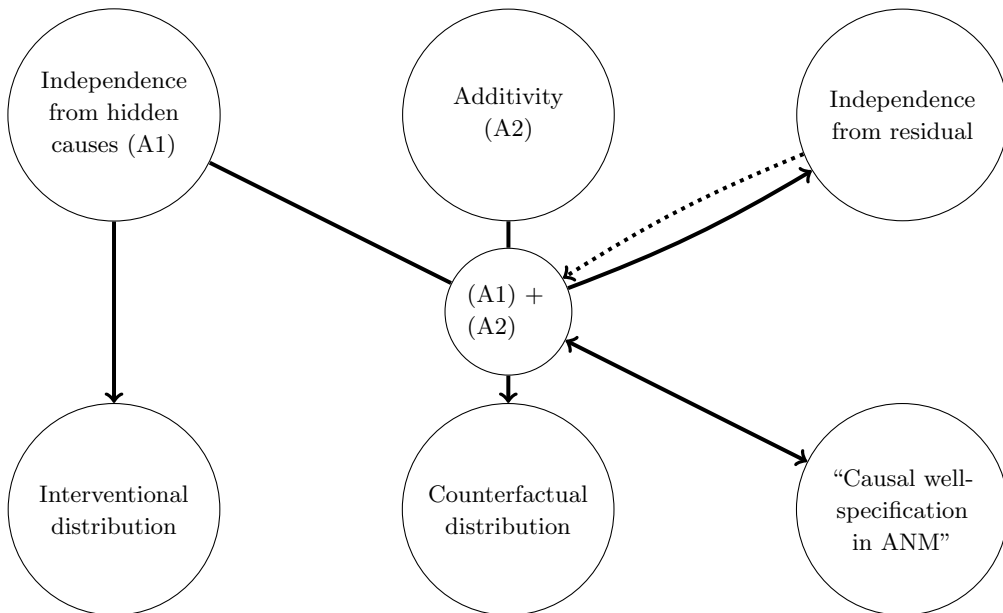
Figure 1: Interconnection between assumptions, causal interpretations, and the testable proxy. Directed edges denote implications, the plus means that both assumptions hold simultaneously, the bidirected edge denotes equivalence per definition, and the dotted edge denotes implication up to pathological cases, i.e., a proxy.

2.3 Global Well-Specification

We recapitulate the ANM assumption for covariates $\mathbf{X}$ and target Y. We call the ANM causally well-specified if

$$Y \leftarrow f_{\mathbf{X}Y}\big(\mathbf{X}_{\mathrm{PA}(Y)}\big) + f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}\big), \quad \text{where} \quad \mathbf{X} \perp \mathbf{H}_{\mathrm{PA}(Y)}. \tag{2}$$

Note that we do not constraint the functional form of $f_{\mathbf{X}Y}(\cdot)$ and $f_{\mathbf{H}Y}(\cdot)$ any further. In particular, the structure does not imply that the functions must be additive in their respective arguments.

The independence condition corresponds to no hidden confounding or hidden mediation. It ensures

$$Y|\mathbf{X} = \mathbf{x} \quad \overset{d}{=} \quad Y|\mathrm{do}(\mathbf{X} \leftarrow \mathbf{x}),$$

where $\overset{d}{=}$ states that two random variables have the same distribution. Assuming faithfulness, it also implies $\mathrm{DE}(Y) \cap M = \emptyset$ since faithfulness ensures $\forall j \in \mathrm{DE}(Y) \; Z_j \not\perp \xi_Y \in \mathbf{H}_{\mathrm{PA}(Y)}$.

The parametrization in (2) is not unique as constants could be moved between the two summands. We let the second have mean 0 such that $\mathbb{E}[Y|\mathbf{X}] = f_{\mathbf{X}Y}\big(\mathbf{X}_{\mathrm{PA}(Y)}\big)$. The additivity condition then ensures that in the counterfactual, where we can change $\mathbf{X}$ without changing any other unobserved noise term, the outcome is exactly shifted by the difference in conditional expectation. Thus, we fully understand the effect of changing $\mathbf{X}$. Denote

point masses at y by δ_y, then,

$$P_Y^{\mathfrak{E}|\mathbf{Z}=\mathbf{z};\mathrm{do}(\mathbf{X}\leftarrow\mathbf{x}')} = \delta_{y'} \quad \text{where} \quad y' = y + \mathbb{E}\big[Y|\mathbf{X} = \mathbf{x}'\big] - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}].$$

The conditions in (2) additionally imply the following global null hypothesis that we aim to check first.

$$H_0: \quad \mathcal{E} \perp \mathbf{X}, \quad \text{where} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}]. \tag{3}$$

Note the subtle difference between (2) and (3). The conditions in (2) are sufficient to fulfill (3) but they are not necessary as such independence could also exist non-causally. A prime example is with jointly Gaussian $\mathbf{Z}$: then, H_0 holds regardless of the independence condition in (2) as $\mathcal{E}$ and $\mathbf{X}$ are uncorrelated. The additivity is always fulfilled since multivariate Gaussianity implies linear additive causal effects. However, except for Gaussian $\mathbf{Z}$ or some other pathological data generating distributions, (3) is a useful proxy for (2), i.e., it allows to check whether $\mathbb{E}[Y|\mathbf{X}]$ represents a true causal effect; see also the discussions on the identifiability of ANM in Hoyer et al. (2008) and Peters et al. (2014).

To test H_0, any valid test for independence of $\mathbf{X}$ and $\mathcal{E}$ can be used.

2.4 Local Well-Specification

If the conditions (2) are partially violated it might still be possible to correctly understand the causal effect for *some* of the predictors $\mathbf{X}_U$ where $U \subseteq \{1, \ldots, p\}$. We say the effect of $\mathbf{X}_U$ is causally well-specified in the ANM with response Y and covariates $\mathbf{X}$ if the following hold.

(A1) The covariates in $\mathbf{X}_{-U}$ form a Markov blanket of $\mathbf{H}_{\mathrm{PA}(Y)}$, i.e., $\mathbf{H}_{\mathrm{PA}(Y)} \perp \mathbf{X}_U|\mathbf{X}_{-U}$.

(A2) $Y \leftarrow f_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\backslash U}\big) + f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\backslash U}\big)$, i.e., the causal effect is additively separable into all terms that include $\mathbf{X}_U$ only with observed $\mathbf{X}$ and all terms that include $\mathbf{H}$ without $\mathbf{X}_U$.

Consider Figure 2 containing two examples of DAGs with hidden parents. On the left, the set $U = \{2\}$ fulfils (A1). On the right, $U = \{1\}$ fulfils it.

Correctness of (A2) depends on the structural assignment for Y. E.g.,
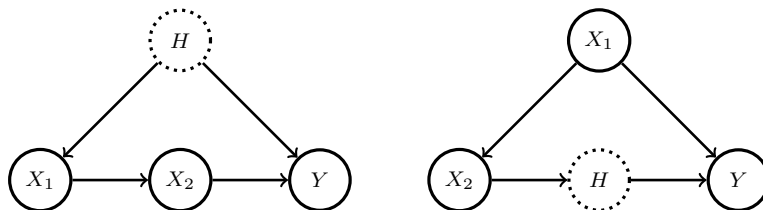
$$Y \leftarrow \sin(X_2) + \sin(H),$$

Figure 2: Left: Structure with a hidden confounder. Right: Structure with a hidden mediator.

is ok for $U = \{2\}$ in the example on the left, but

$$Y \leftarrow \sin(X_2 + H),$$

is not.

If $\exists k \in M \cap \mathrm{DE}(Y)$, i.e., we observe one or more descendants of Y, all X_j where $j \in \mathrm{PA}(Y)$ are, up to faithfulness, in any Markov blanket of $\xi_Y \in \mathbf{H}_{\mathrm{PA}(Y)}$ since they have a common child with respect to the measured covariates. Thus, sets containing measured parents do not fulfill (A1). When choosing (causal) predictors, one aims for $M \cap \mathrm{DE}(Y) = \emptyset$, but there is in general no guarantee for it. So, this is another potential model misspecification. Violations of faithfulness are irrelevant for (A1): if there are effects from $\mathbf{H}_{\mathrm{PA}(Y)}$ to $\mathbf{X}_U$ or vice-versa that cancel out each other, we receive the same implications as if there were no such effects unless (B1) is violated; see below.

We impose no constraints onto $f_{\mathbf{X}_U Y}(\cdot)$ and $f_{\mathbf{H}Y}(\cdot)$ in (A2). Hence, the first summand could be zero, which is the case for $U \cap \mathrm{PA}(Y) = \emptyset$.

(A1) ensures that

$$Y | \mathbf{X}_U = \mathbf{x}_U, \mathbf{X}_{-U} = \mathbf{x}_{-U} \quad \overset{d}{=} \quad Y | \mathrm{do}(\mathbf{X}_U \leftarrow \mathbf{x}_U), \mathbf{X}_{-U} = \mathbf{x}_{-U}$$

whenever both are defined. This follows from the second rule of do-calculus (Pearl, 2012). After removing edges out of $\mathbf{X}_U$, dependence between $\mathbf{X}_U$ and Y could only be induced by a common ancestor or a path from Y to $\mathbf{X}_U$. But, these are all blocked by $\mathbf{X}_{-U}$, on which we condition, by the assumption.

Combined with (A2), we get two implications under an additional technical assumption.

(B1) Let $\{A, B, C\}$ be disjoint subsets of $\{1, \ldots, q\}$ in model (1). Then,

$$\mathbf{Z}_A \perp \mathbf{Z}_B | \mathbf{Z}_C \implies \mathbf{Z}_A | \mathbf{Z}_B = \mathbf{z}_B, \mathbf{Z}_C = \mathbf{z}_C \overset{d}{=} \mathbf{Z}_A | \mathbf{Z}_B = \mathbf{z}'_B, \mathbf{Z}_C = \mathbf{z}_C \quad \forall \mathbf{z}_B, \mathbf{z}'_B, \mathbf{z}_C.$$

This means that there are no unobservable dependencies on null sets of the observational distribution which is natural to assume except for pathological data. In general, independence only implies the latter equality for almost all $\mathbf{z}_B, \mathbf{z}'_B, \mathbf{z}_C$. A counterexample to (B1) would be the following SCM with continuous and univariate components Z_A, Z_B, and Z_C

$$
\begin{aligned}
Z_C &\leftarrow \xi_C, \\
Z_B &\leftarrow Z_C + \xi_B, \\
Z_A &\leftarrow Z_C + \xi_A + \mathbb{1}_{\{Z_B = 0\}}.
\end{aligned}
$$

Then,

$$Z_A \perp Z_B | Z_C, \quad \text{but, e.g.,} \quad Z_A | Z_B = 1, Z_C = 0 \overset{d}{\neq} Z_A | Z_B = 0, Z_C = 0,$$

i.e., there is no observable conditional dependence between Z_A and Z_B, but we could, in theory, create an intervention that provokes unexpected behaviour. Therefore, we exclude such hidden dependencies.

Theorem 1. Assume the model (1) with (B1). Let $\mathbf{X}_U$ be a set of covariates fulfilling (A1) and (A2), then

$$P_Y^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\mathrm{do}\left(\mathbf{X}_U\leftarrow\mathbf{x}'_U,\mathbf{X}_{-U}\leftarrow\mathbf{x}_{-U}\right)} = \delta_{y'} \quad \text{where}$$
$$y' = y + \mathbb{E}\left[Y|\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U}\right] - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

for $(\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U})$ in the support of the observational distribution. Further, $H_{0,U}$ holds, where

$$H_{0,U}: \quad \mathcal{E} \perp \mathbf{X}_U|\mathbf{X}_{-U}, \quad \text{with} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}]. \tag{4}$$

The first implication means that in the counterfactual, where we can change $\mathbf{X}_U$ without changing $\mathbf{X}_{-U}$ or ξ_k in (1) $\forall k \notin M$, the effect on Y is fully determined by the shift in conditional expectation. Thus, we understand the causal effect of this theoretical intervention. Note that with (A1) and (B1), not changing $\mathbf{X}_{-U}$ and ξ_k $\forall k \notin M$ is equivalent to not changing $\mathbf{X}_{-U}$ and $\mathbf{H}$, i.e., all other variables apart from $\mathbf{X}_U$ and Y remain unchanged; see also the proof in Appendix A.1. More generally, including cases where $(\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U})$ is outside the support of the observational distribution, one could replace

$$\mathbb{E}\left[Y|\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U}\right] \quad \text{by} \quad \mathbb{E}\left[Y|\mathrm{do}\left(\mathbf{X}_U \leftarrow \mathbf{x}'_U\right), \mathbf{X}_{-U} = \mathbf{x}_{-U}\right]$$

which are equivalent if both are defined as discussed above. However, this is not estimable outside the data support. $H_{0,U}$ (4) is equivalent to saying that $\{1, \dots p\} \setminus U$ defines a Markov blanket of the residual $\mathcal{E}$.

We note that the implication of (A1) would be of practical interest on its own. However, as we do not know of any useful proxy for it that can be calculated by the observational distribution, we always consider the combination of (A1) and (A2) as the object of interest.

The local null hypothesis $H_{0,U}$, which can be checked by the observational distribution alone, serves as a proxy for (A1) and (A2). Again, a multivariate Gaussian distribution is an example where (4) holds regardless of (A1). However, for other data generating distributions, we consider (4) to be a good proxy to see whether (A1) and (A2) might hold.

Of most interest are the sets

$$W \in \underset{U:H_{0,U} \text{ is true}}{\arg\max} |U|. \tag{5}$$

As W is of maximum size, $\{1, \dots p\} \setminus W$ is of minimum size. Thus, it is not only a Markov blanket but a Markov boundary of $\mathcal{E}$ such that uniqueness of W is implied by the uniqueness of the Markov boundary. This is guaranteed if the so-called intersection property holds (Pearl, 1988, Chapter 3).

(B2) $\mathcal{E} \perp \mathbf{X}_A|\mathbf{X}_B, \mathbf{X}_C \quad \text{and} \quad \mathcal{E} \perp \mathbf{X}_B|\mathbf{X}_A, \mathbf{X}_C \implies \mathcal{E} \perp \mathbf{X}_A, \mathbf{X}_B|\mathbf{X}_C$
for any partition A, B, C of $\{1, \dots, p\}$.

$\mathbf{X}$ having full support with respect to the product of the domains of the individual X_j is sufficient for the intersection property and hence uniqueness of the Markov boundary. Strictly weaker, necessary and sufficient conditions are discussed by Peters (2015).

One estimation strategy would be to consider the individual hypothesis and output the collection of all variables for which these individual hypotheses are true

$$H_{0,j}: \quad \mathcal{E} \perp X_j | \mathbf{X}_{-j}, \quad \text{where} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}]$$
$$\tilde{W} = \{j : H_{0,j} \text{ is true}\}. \tag{6}$$

We can relate this to the Markov boundary.

Theorem 2. Assume the model (1). Let W be any set as in (5) and $\tilde{W}$ as in (6). Then, $W \subseteq \tilde{W}$. If the intersection property (B2) holds, $W = \tilde{W}$, and W is unique.

In general, (A1) and (A2) do not imply that the ANM (2) with only $\mathbf{X}_U$ as predictors is causally well-specified. Therefore, this set cannot be found by looping over all subsets of $\mathbf{X}$ and testing (3).

Recall the examples in Figure 2. In the left structure, $W = \{2\}$ if (A2) holds for X_2. But, $X_2 \to Y$ is not a causally well-specified ANM unless faithfulness is violated, i.e., $X_2 \perp H$. Similarly, on the right, it holds $W = \{1\}$ if (A2) holds for X_1. But, $X_1 \to Y$ is not a causally well-specified ANM unless faithfulness is violated, i.e., $X_1 \perp H$.

Note also that there is an unobserved causal path (Maeda and Shimizu, 2021) from X_1 to Y. This means that their method which exploits different potential parental sets to obtain independent residuals cannot identify this edge. Nevertheless, the edge can be characterized as causally well-specified when considering conditional independence criteria.

We emphasize that the characterizations in this Section 2 provide the fundamental basis to define the concepts of global and local causal well-specification. This then enables the construction of algorithms that aim to estimate causal well-specification based on finite sample observational data, as discussed next.

3. Estimating the Set of Well-Specified Predictor Variables

We subsequently focus on a specific method to assess conditional dependence. Of course, different estimators could be used as well. The intuition of how conditional independence relates to causal well-specification stays the same. The practical algorithm to estimate the set of variables with well-specified effect is given in Section 3.3.

Throughout this section, we assume that we have n i.i.d. observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $y_1, \ldots, y_n$ of $\mathbf{X}$ and Y respectively. More compactly, this data can be written as $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Also, define the unobserved $\epsilon_i = y_i - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}_i]$.

3.1 Making Use of FOCI (Feature Ordering by Conditional Independence)

One estimation strategy would be to test the hypotheses in (6) for all j. Conditional independence testing is a hard problem on its own; see, e.g., Shah and Peters (2020). Here, it is even more challenging as we need to rely on estimated residuals rather than the error terms directly. Instead of testing, we use FOCI (Feature Ordering by Conditional Independence) by Azadkia and Chatterjee (2021). This method estimates a Markov blanket, which they call a sufficient set, of a target variable, and they give guarantees which hold

with high probability for large enough sample size. Thus, it can find a superset of the Markov boundary of $\mathcal{E}$, say, $\hat{S}$, such that $\left(\{1, \ldots, p\} \setminus \hat{S}\right) \subseteq W$.

Before reviewing the most important concepts of FOCI, we emphasize what our contribution to the subsequent results is. Here, we need to deal with the harder problem of applying FOCI to the estimated residuals $\hat{\boldsymbol{\epsilon}}$ instead of the true, unobserved residuals $\boldsymbol{\epsilon}$. We extend the theory from Azadkia et al. (2021) to this case and provide asymptotic guarantees in Section 3.2. As additional assumptions, we require only a weak form of consistency for the regression estimates as well as continuous residuals. An example demonstrating the pitfalls of discrete residuals is given in Section 3.2.1. Furthermore, we show a new result for transforming the data before applying FOCI; see Proposition 2. Finally, we suggest an algorithm that yields more stable estimates in Section 3.3.

We provide now some background on FOCI by focusing on its main concepts. The precise definitions can be found in Appendix A.3. Assume we want to consider if

$$\mathcal{E} \perp \mathbf{X}_U | \mathbf{X}_S.$$

Azadkia et al. (2021) define a coefficient of conditional dependence, $T(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) \in [0, 1]$, which is 0 for conditional dependence, 1 if $\mathcal{E}$ is almost surely a function of $\mathbf{X}_U$ given $\mathbf{X}_S$, and in between otherwise. A slightly different coefficient with the same properties is defined for empty conditioning sets. $T(\cdot)$ can be decomposed as

$$T(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) = \frac{Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S)}{S(\mathcal{E}, \mathbf{X}_S)},$$

with a nonnegative numerator and denominator. Thus, conditional independence is also equivalent to $Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) = 0$. By construction

$$Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) = Q(\mathcal{E}, \{\mathbf{X}_U, \mathbf{X}_S\}) - Q(\mathcal{E}, \mathbf{X}_S)$$

using the version of $Q(\cdot)$ without a conditioning set. FOCI is designed to greedily find an additional covariate that maximizes $T(\mathcal{E}, X_j | \mathbf{X}_S)$ assuming $\mathbf{X}_S$ has already been chosen. This is equivalent to greedily maximizing $Q(\mathcal{E}, \{X_j, \mathbf{X}_S\})$ as the normalization $S(\cdot)$ above does not depend on the candidate variable. For practical evaluation, one can use a sample estimate $Q_n(\boldsymbol{\epsilon}, \mathbf{x}_S)$. This only depends on the relative order of $\boldsymbol{\epsilon}$ and is not guaranteed to be non-decreasing when adding additional covariates to S. Hence, FOCI stops when no candidate variable yields and improvement in $Q_n(\cdot)$, i.e.,

$$\forall j \in \{1, \ldots, p\} \setminus S : \quad Q_n(\boldsymbol{\epsilon}, \{\mathbf{x}_j, \mathbf{x}_S\}) \leq Q_n(\boldsymbol{\epsilon}, \mathbf{x}_S).$$

For large enough data, $Q_n(\boldsymbol{\epsilon}, \{\mathbf{x}_j, \mathbf{x}_S\}) \approx Q(\mathcal{E}, \{X_j, \mathbf{X}_S\})$ such that the algorithm does not stop before the estimated set $\hat{S}$ is a Markov blanket, i.e., it includes all necessary covariates $\{1, \ldots, p\} \setminus W$ with high probability. However, there is in general no guarantee against superfluous inclusion to $\hat{S}$ and we get $\left(\{1, \ldots, p\} \setminus \hat{S}\right) \subseteq W$.

For power purposes, it can be advantageous to consider a certain non-monotonic transformation $g(\mathcal{E})$ as input to FOCI. Intuitively, $T(g(\mathcal{E}), X_j | \mathbf{X}_S)$ measures nonparametrically how much X_j increases the explicative power for $g(\mathcal{E})$. Hence, transforming $\mathcal{E}$ such that this relative explicative power increases, makes detection easier. In particular, we suggest

the absolute value function. For this, we provide a precise result for symmetric data below. Although exact symmetry is hardly the case except for toy examples, the intuition is that the dependence of $\mathcal{E}$ on $\mathbf{X}$ can be mainly in the second moment, i.e., the scale. Hence, the absolute value transform is then beneficial. For our general results, we assume that $g(\cdot)$ is an l-Lipschitz function whose level sets have Lebesgue measure 0.

For the precise definitions for $T(\cdot)$ and $Q(\cdot)$; see (2.1) and (11.1) in Azadkia and Chatterjee (2021) or Appendix A.3 here. FOCI greedily increases the set of predictors to maximize Q.

Proposition 1. Let $S \subseteq \{1, \ldots, p\}$. If $\mathcal{E}|\mathbf{X}_S$ has a continuous and symmetric (around 0) distribution, it holds

$$T(|\mathcal{E}|, \mathbf{X}_S) = 4T(\mathcal{E}, \mathbf{X}_S) \quad \text{and} \quad Q(|\mathcal{E}|, \mathbf{X}_S) = 4Q(\mathcal{E}, \mathbf{X}_S).$$

These larger population values can improve the algorithm's performance.

In general, we require some sort of consistency for our regression estimates and our discussion allows any reasonable choice of regression (machine learning) techniques. While as in the population case rejections of the null hypothesis could only be due to hidden confounding or additively non-separable functions, one must always consider insufficient explicative power of the applied regression (machine learning) method as a further reason in the finite sample case.

We consider two different algorithms. For notational simplicity, we call the data that is input to FOCI $(\hat{\boldsymbol{\epsilon}}, \mathbf{x})$ in our theoretical derivations regardless of the applied algorithm, i.e., we omit the superscript in the splitting case. The advantage of Algorithm 2 is that the residuals estimated on the hold-out split are still i.i.d. which simplifies things, at least analytically. Furthermore, the sample splitting idea enables further favourable algorithms to be presented in Section 3.3.

Algorithm 1 In-sample FOCI

    Input i.i.d. data $\mathbf{x} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$, and function $g(\cdot)$
    Output estimated set of variables $\hat{S}$ for which null hypothesis (4) is rejected
1: Get an estimate $\hat{f}(\mathbf{X})$ for $\mathbb{E}[Y|\mathbf{X}]$ using a certain regressor
2: Estimate the residuals as $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{f}(\mathbf{x})$
3: Apply FOCI (Azadkia and Chatterjee, 2021) to the data $(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x})$ to get the set $\hat{S}$

Algorithm 2 Sample splitting FOCI

    Input i.i.d. data $\mathbf{x} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$, and function $g(\cdot)$
    Output estimated set of variables $\hat{S}$ for which null hypothesis (4) is rejected
1: Split the data uniformly at random into two disjoint parts of sizes $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$, say, $\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right)$ and $\left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}\right)$
2: Get an estimate $\hat{f}(\mathbf{X})$ for $\mathbb{E}[Y|\mathbf{X}]$ using a certain regressor on the data $\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right)$
3: Estimate the residuals as $\hat{\boldsymbol{\epsilon}}^{(2)} = \mathbf{y}^{(2)} - \hat{f}\left(\mathbf{x}^{(2)}\right)$
4: Apply FOCI (Azadkia and Chatterjee, 2021) to the data $\left(g\left(\hat{\boldsymbol{\epsilon}}^{(2)}\right), \mathbf{x}^{(2)}\right)$ to get the set $\hat{S}$

3.2 Asymptotic Results

We generally make the following assumptions for applying FOCI to an estimated $\hat{\boldsymbol{\epsilon}}$.

(B3) $|\hat{\epsilon}_i - \epsilon_i| = \mathcal{O}_p(1)$.

(B4) $\mathcal{E}$ is a continuous random variable.

(B5) $\nexists S \subseteq \{1, \ldots, p\}$ such that $\mathbf{X}_{-S} \not\perp \mathcal{E}|\mathbf{X}_S$ but $\mathbf{X}_{-S} \perp g(\mathcal{E})|\mathbf{X}_S$.

The probability in (B3) is with respect to both the regression estimate and the new data point i. The assumption is slightly different depending on which algorithm is applied. Apart from invoking (B4) for the proofs, we provide a simple example in Section 3.2.1 to show that discrete distributions can lead to inconsistency.

The main proof ingredient for adapting the results to our setting is showing that for random indices i and l the probability that the estimated residuals imply a different relative ordering than the true residuals approaches 0.

With sample splitting, we obtain a consistency result analogous to Azadkia and Chatterjee (2021). We require the same regularity conditions as they do. Let

$$\delta = \min_{j,S:\ T(g(\mathcal{E}), X_j|\mathbf{X}_S) > 0} Q(g(\mathcal{E}), \{\mathbf{X_S}, X_j\}) - Q(g(\mathcal{E}), \mathbf{X_S}),$$

i.e., the lowest difference in $Q(\cdot)$ we should be able to detect.

(A1') There are nonnegative real numbers β and C such that for any set $S \subseteq \{1, \ldots, p\}$ of size $s \leq 1/\delta + 2$, any $\mathbf{x}_S, \mathbf{x}'_S \in \mathbb{R}^s$ and any $t \in \mathbb{R}$,

$$\left| P(g(\mathcal{E}) \geq t|\mathbf{X} = \mathbf{x}_S) - P(g(\mathcal{E}) \geq t|\mathbf{X} = \mathbf{x}'_S) \right|$$
$$\leq C(1 + \|\mathbf{x}_S\|^{\beta} + \|\mathbf{x}'_S\|^{\beta})\|\mathbf{x}_S - \mathbf{x}'_S\|.$$

(A2') There are positive numbers C_1 and C_2 such that for any S of size $s \leq 1/\delta + 2$ and any $t > 0$, $\mathbb{P}(\|\mathbf{X}_S\| \geq t) \leq C_1 e^{-C_2 t}$.

Theorem 3. Suppose that the regularity assumptions (A1') and (A2') (Azadkia and Chatterjee, 2021) for the data $(g(\mathcal{E}), \mathbf{X})$ hold as well as conditions (B2) - (B5). Let $\hat{S}$ be the output of Algorithm 2. There are positive real numbers L_1, L_2 and L_3 that do not depend on the sample size such that

$$\mathbb{P}\left(\hat{S} \supseteq \{1, \ldots, p\} \setminus W\right) \geq 1 - L_1 p^{L_2} \exp(-L_3 n).$$

Without sample splitting, $(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)$ are not independent copies. Therefore, the bounded difference inequality (McDiarmid et al., 1989) which is applied to obtain the exponential probability decay cannot be used. Nevertheless, convergence in probability is still true.

Theorem 4. Assume the conditions of Theorem 3. Let $\hat{S}$ be the output of Algorithm 1. Then,

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{S} \supseteq \{1, \ldots, p\} \setminus W\right) = 1.$$

This result is derived by a simple application of the Markov inequality instead of the bounded difference inequality. As the $(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)$ become decreasingly dependent from another with increasing sample size, we conjecture that the true convergence rate could be similar to the one for sample splitting.

3.2.1 Discrete $\mathcal{E}$

For our results, we invoked assumption (B4), i.e., the residuum is a continuous random variable. A simple toy example shows that a discrete random variable might invalidate the asymptotic guarantees. Use the definition (2.1) in Azadkia and Chatterjee (2021) for $T(\cdot)$, i.e., $T(\mathcal{E}, X) = 0$ if and only if $\mathcal{E}$ and X are independent. Let $T_n(\cdot)$ be its suggested sample estimate; see also Appendix A.3.

Proposition 2. Let X be a bounded, continuous random variable and $\mathcal{E}$ a centered random variable that is uniformly distributed over a discrete set of size $k > 1$ independent from X such that $T(\mathcal{E}, X) = 0$. Let $Y \leftarrow X\beta + \mathcal{E}$ for some $\beta \neq 0$. Apply linear least squares regression, which fulfils (B3), to n i.i.d. copies $(\mathbf{y}, \mathbf{x})$ to get the estimates $\hat{\boldsymbol{\epsilon}}$. It holds

$$\mathbb{E}[T_n(\hat{\boldsymbol{\epsilon}}, \mathbf{x})] \overset{n \to \infty}{\to} \frac{1}{k^2} > 0.$$

We provide some intuition while the detailed proof can be found in Appendix A.7. As $\mathbf{x}$ is continuous, it is never perfectly orthogonal to $\boldsymbol{\epsilon}$, and the least squares estimator is slightly off. Then, within each of the k groups the ranking of the $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \mathbf{x}\left(\beta - \hat{\beta}\right)$ is exactly according to the ranking of the $\mathbf{x}$ or inverted such that there is some non-vanishing dependence that FOCI detects.

3.3 Practical Algorithm

Although we can consistently find a Markov blanket (but not necessarily the minimal Markov boundary) using Algorithms 1 or 2 as the sample size grows, there are several drawbacks to that. First, there is no protection against including superfluous variables into $\hat{S}$ and typically this happens with non-negligible probability. Second, for low sample sizes, $\hat{S}$ can miss out on some variables.

To partially remedy these issues, we incorporate ideas from multisplitting (Meinshausen et al., 2009) and stability selection (Meinshausen and Bühlmann, 2010). We apply Algorithm 2 repeatedly with several random data partitions. Inspired by Shah and Samworth (2013) who suggest using "complementary pairs", i.e., both halves of every split, we let each halve be used once for estimating the conditional mean and once for independence testing.

As unconditional independence is easier to assess than conditional independence, we first test for H_0 as in (3). For this, we apply the test by Pfister et al. (2018). The case where only estimates of the residuals are available is explicitly discussed in their work. Then, we combine the p-values over the different splits as suggested by Meinshausen et al. (2009). Only if the global model is rejected, the individual covariates are inspected.

If we cannot trust the overall model, we only consider the effects of variables that are selected substantially less than others by the FOCI algorithm to be causally well-specified. We split the variables into two groups: those that are selected by FOCI below average over the splits and the others. For the latter group, we reject $H_{0,j}$. Each variable from the first group we compare to the least selected variable from the second group with some proportion test such as Fisher's exact test. The variables that show significant differences are added to the estimated well-specified set $\hat{W}$. Notably, there is no exact interpretation of the significance level used for these tests, but the intuition that a lower significance level

leads to fewer false positives in the set $\hat{W}$ remains true. In contrast, a lower significance level for the preceding test of the global model leads to the methods becoming more liberal.

The intuition behind splitting at the mean is the following. For large enough sample size, the necessary variables are selected by FOCI in almost every split, see Theorem 3, while the variables with causally well-specified effect could be selected with some probability much lower than 1. The mean separates the two groups and there is a significant difference in the selection fraction of the two groups. For a low sample size, the behaviour of FOCI is more random. However, as long as no variables stand out, we do not add any to $\hat{W}$, i.e., if $H_{0,j}$ is not true, the probability $\mathbb{P}\left(j \in \hat{W}\right)$ is moderately low. However, it is lower bound by the type II error of the global test. This is fundamental to our idea. If the sample size is such that the global test, i.e., unconditional independence testing, does not work well yet, the local analysis is also not of much use. We summarize the procedure in Algorithm 3.

Algorithm 3 Selection of variables with well-specified effect using multiple splits

Input i.i.d. data $\mathbf{x} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$, function $g(\cdot)$, number of repetitions B, and significance levels α and $\tilde{\alpha}$.

Output estimated set of variables $\hat{W}$ with causally well-specified effect

1: $n_j = 0 \ \forall j = 1, \ldots, p$
2: **for** $b = 1$ to B **do**
3:     Split the data uniformly at random into two disjoint parts of sizes $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$, say, $\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right)$ and $\left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}\right)$
4:     Get an estimate $\hat{f}(\mathbf{X})$ for $\mathbb{E}[Y|\mathbf{X}]$ using a certain regressor on the data $\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right)$
5:     Estimate the residuals as $\hat{\boldsymbol{\epsilon}}^{(2)} = \mathbf{y}^{(2)} - \hat{f}(\mathbf{x}^{(2)})$
6:     Apply the HSIC test to the data $\left(\hat{\boldsymbol{\epsilon}}^{(2)}, \mathbf{x}^{(2)}\right)$ to get the p-value p^b.
7:     Apply FOCI (Azadkia and Chatterjee, 2021) to the data $\left(g\left(\hat{\boldsymbol{\epsilon}}^{(2)}\right), \mathbf{x}^{(2)}\right)$ to get the set $\hat{S}^b$
8:     Swap the roles of $\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right)$ and $\left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}\right)$ and repeat the previous steps to get p^{B+b} and $\hat{S}^{B+b}$
9: Combine the p-values $p^1, \ldots p^{2B}$ (Meinshausen et al., 2009) and get the model p-value p^0.
10: **if** $p^0 > \alpha$ **then**
11:     $\hat{W} = \{1, \ldots, p\}$
12: **else**
13:     $\hat{W} = \emptyset$
14:     **for** $j = 1$ to p **do**
15:         $n_j = \sum_{b=1}^{2B} j \in \hat{S}^b$
16:     $\bar{n} = \sum_{j=1}^{p} n_j/p$
17:     $n^{\min} = \min_{j : n_j \geq \bar{n}} n_j$
18:     **for** $j = 1$ to p **do**
19:         **if** $n_j < \bar{n}$ and $proportion.test\left(n_j, n^{min}, 2B\right) \leq \tilde{\alpha}$ **then**
20:             $\hat{W} = \hat{W} \cup j$

---

4. Simulation Example

We evaluate the method on a simple SCM represented by the DAG in Figure 3. We let the causal effects be non-monotonic functions. As discussed in Schultheiss and Bühlmann (2023), non-monotonic effects can lead to stronger dependence between residual and predictor in the wrong direction. Hence, we are a bit more sample-efficient than with monotonic functions. The effects have the following form

$$f(X_j) = \alpha_1 |X_j|^{\beta_1} \text{sign}(X_j) + \alpha_2 |X_j|^{\beta_2},$$

where the parameters are randomly sampled and differ for every simulation run. The causal effect on Y is additive in the parents. We standardize and normalize the effects. The additive error terms are either normal, uniform, or Laplace with variance 1 for the root nodes and $1/4$ for the others. The different distributions are randomly assigned to the different nodes; two of each.

We consider all possible subsets of size 3 as observed predictors. Denote this observed subset by M. For $M = \{1, 2, 3\}$ and $M = \{1, 3, 5\}$ the additive noise model is causally well-specified.

We consider 100 different random setups for sample sizes 10^2 to 10^5. For each, we consider all possible M. To get $\hat{W}$ we apply Algorithm 3 with $B = 25$ splits and the absolute value function as $g(\cdot)$.

For the regression, we apply eXtreme Gradient Boosting implemented in the R-package `xgboost` (Chen et al., 2021). Other regression (machine learning) techniques could be used instead if they are flexible enough. We fix our choice here for this proof of concept. We use the respective left-out split of the data for early stopping when fitting the regression functions. This is a slight violation of our theoretical algorithm where the residuals are perfectly independent. We use the authors' implementation of FOCI (Azadkia et al., 2021) and dHSIC (Pfister and Peters, 2019).

For the predictor sets where global causal well-specification does not hold, we consider the false positive rate (FPR) $\hat{\mathbb{P}}\left(j \in \hat{W} | j \notin W\right)$ and the true positive rate (TPR) $\hat{\mathbb{P}}\left(j \in \hat{W} | j \in W\right)$ for adding predictors to the set $\hat{W}$. Here, $\hat{\mathbb{P}}(\cdot)$ denotes the empirical probability over our simulation runs. We fix $\alpha = 0.05$ and consider varying values of $\tilde{\alpha}$. The resulting rates are on the left in Figure 4.

For $n = 10^2$, a low FPR is not attainable because the p-values for (3) are not reliably small enough, and Algorithm 3 often terminates before considering the individual covariates. However, even for this low sample size, we get a performance that is clearly better than
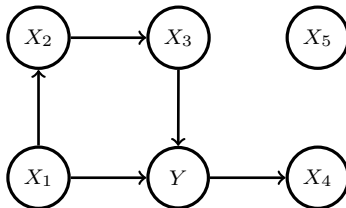

Figure 3: DAG representing the SCM in the simulation.

random guessing. For large sample sizes, the FPR becomes very low which is in agreement with Theorem 3. The lack of power is mainly due to the subsets of predictors with $|W| > 1$. FOCI chooses superfluous covariates with non-vanishing probability for every sample size. Hence, the two covariates with causally well-specified effects may be selected with a frequency that differs a lot between the two. If one then appears to be more similar to the covariate with not well-specified effect, our algorithm misses out on this such that $\hat{W} \subset W$.

For comparison, we also show the results if we instead only consider a single random split where 50% of the data is used to estimate the residuals and the other 50% to assess independence. If H_0 is rejected we apply Algorithm 2 (using the same splits) and choose $\hat{W}$ to be the complement of the set chosen by FOCI. Except for $n = 10^2$, this lies below the curve for multiple splits, i.e., there is an $\tilde{\alpha}$ that is better in terms of both FPR and TPR. Further, our default choice $\tilde{\alpha} = 0.01$ is more conservative. For large enough sample size, using $\tilde{\alpha} = 0.01$ leads to more power than considering a single split. Hence, even though the problem is hard in general, aggregating information over multiple random splits of the same data set can lead to a performance boost.

We also evaluate the testing of H_0 (3). For this, we show the empirical cumulative distribution function of the obtained p-values in the middle of Figure 4. We consider the p-value aggregated over the splits as well as the individual p-values considering single splits. For the largest sample sizes, the distribution of both is visibly not distinguishable from a point mass at 0. We omit this in the plot for the sake of overview. For $n = 10^2$ and $n = 10^3$, aggregating the p-values over splits helps to reject the global model for most possible significance levels. The acceptance rate for the global model poses a lower bound to the attainable FPR for every subsequent per-covariate analysis. For $n = 10^2$ and $\alpha = 0.05$,
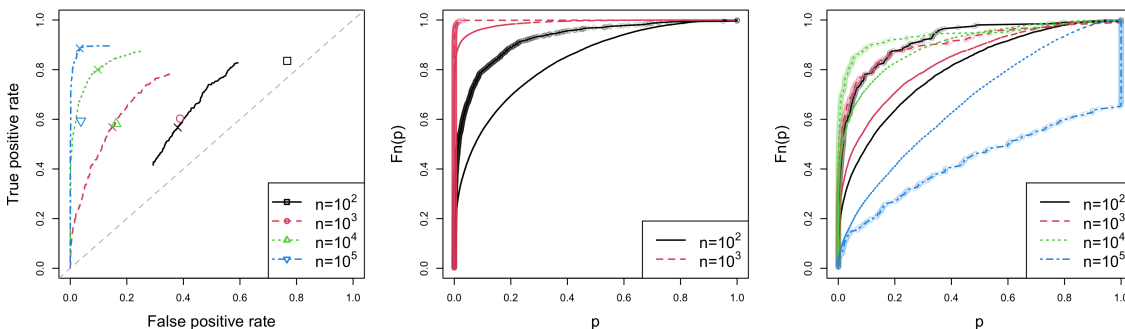

Figure 4: The results are based on 100 simulation runs. On the left: False positive rate versus true positive rate obtained with Algorithm 3 for varying $\tilde{\alpha}$ and $\alpha = 0.05$. The crosses correspond to $\tilde{\alpha} = 0.01$. The other symbols describe the performance of Algorithm 2. In the middle: empirical cumulative distribution function of the p-values obtained with HSIC. We compare the raw p-values from each split (lines only) to the cumulated p-value per simulation run (lines with dots). On the right: the same for the models fulfilling H_0.

17

this rate is around 0.56 for single splits and reduced to roughly 0.33 by aggregating. This confirms the usefulness of the multisplitting idea.

We also consider the distribution of the p-values for the two subsets of predictors that yield causally well-specified models. This is shown on the right side of Figure 4. We see that the raw p-values are too liberal. By construction, this effect is enhanced by aggregation over the splits. For increasing sample size, there are two competing effects. The regression approximation becomes better leading to less dependent residuals. But, the tests become more powerful in detecting spurious dependence. As the HSIC implementation cannot handle $5 \cdot 10^4$ samples, we only test with 10^4 samples per split. Hence, the p-values for $n = 10^5$ are likely more liberal theoretically. In summary, we see that testing for (3) is already difficult per se. However, one can also see it the other way around: if the regression is unable to render the residuals independent one should not trust the obtained function even if there was a true underlying ANM.

In this example, fitting only additive functions with no interactions between the measured covariates leads to the same conclusion given perfect regression fit and independence tests since the data follow a CAM (Bühlmann et al., 2014). Hence, if one restricts the analysis to additive functions due to pre-knowledge or just by assumption the problem could become easier. When applying GAM regression as implemented in `mgcv` (Wood, 2011), the results for the causally not well-specified predictor sets remain qualitatively similar. The p-values for the models fulfilling H_0 are still visibly clearly not uniformly distributed. But, they become less liberal. This is as finding the true conditional mean and hence the true independent residuals becomes easier. For $n = 10^5$ the distribution of the raw p-values is sufficiently close to uniform such that the aggregated p-values are even super-uniform. Again, this needs to be taken with a grain of salt as not all samples can be used for testing independence.

5. Real Data Analysis

We consider the K562 data set provided by Replogle et al. (2022). We follow the preprocessing in the benchmark of Chevalley et al. (2023). Then, the data set contains 162,751 measurements of the activity of 622 genes: 10,691 of the measurements are taken in a purely observational environment while the remaining are obtained under various interventions. For each gene, there exists an environment where it has been intervened on by a knockdown using CRISPRi (Larson et al., 2013), i.e., its activity is reduced. As our method is designed for i.i.d. data, we only consider the observational environment henceforth. With the interventions, some sanity checks of our findings are possible as discussed below.

We make a pre-selection of the measured covariates before applying our method. There are 28 genes that are active, i.e., greater than 0, in each measurement in the observational sample. We restrict our analysis to these and call them X_1 to X_{28} for simplicity. Within these 28, we estimate Markov blankets using FOCI. For each of the 28 genes, two estimates are implied: all the genes selected by FOCI when this covariate is the target as well as all the genes for which this covariate is in the output of FOCI. As the target, say Y, we choose the one with the highest agreement between the two estimated sets in terms of intersection size relative to the size of the union. For the target Y, we then consider the intersection of

the Markov blankets mentioned above (where Y is the target or appears in the output of FOCI). This results in three predictors, X_{10}, X_{12}, and X_{15}.

With the selected target and predictors we run Algorithm 3 with $B = 25$ splits using `xgboost` for regression. There is a strong indication against the global null hypothesis (3) with a p-value of roughly 10^{-27}. Hence, we proceed to the per-covariate analysis. Covariate X_{15} is in $\hat{S}^b$ 41 out of 50 times while as for the others it is only 22 (X_{10}) and 19 (X_{12}). Hence, the effects of the latter appear to be causally well-specified and we get the set $\hat{W} = \{X_{10}, X_{15}\}$ when running Algorithm 3 with our suggested default of $\tilde{\alpha} = 0.01$.

To assess the success of our method, we now consider the available interventional data. Comparing the distribution of X_k when the activity of X_j is reduced by an external intervention to its observational distribution, gives an assessment of whether there is a causal effect from X_j to X_k. We do this using a Mann-Whitney U test. Intervening on any of the three predictor covariates appears to highly influence the activity of Y with p-values of the order 10^{-4}, 10^{-13}, and 10^{-6}. In the reverse direction, intervening on Y does not have strong influence on X_{10} ($p \approx 0.1$) and X_{12} ($p \approx 0.5$) but on X_{15} ($p \approx 4 * 10^{-5}$). Thus, there appears to be some cyclic effect between Y and X_{15}. Hence, it is less appropriate to consider its regression effect to be causally well-specified whereas our estimated well-specification for X_{10}, X_{12} on Y is compatible with the validation analysis based on interventional data.

Finally, we can also compare how well our regression model trained on the observational data performs on data from the different interventional environments. We do this comparison in terms of absolute bias relative to Y's mean activity in the observational sample, i.e.,

$$RB^{X_j \to Y} = \frac{\left|\sum_{i \in \mathcal{D}_j} y_i - \hat{f}(\mathbf{x}_i)\right|/|\mathcal{D}_j|}{\sum_{i \in \mathcal{D}_{\mathcal{O}}} y_i/|\mathcal{D}_{\mathcal{O}}|}, \tag{7}$$

where $\mathcal{D}_j$ denotes the data points where X_j is knocked down, $\mathcal{D}_{\mathcal{O}}$ the observational data, and $\hat{f}(\cdot)$ is trained on $\mathcal{D}_{\mathcal{O}}$. This suggests that generalization to the environment where a knockdown is applied to X_{15} works the least with a relative bias (7) of about 12% while in the other environments it is roughly 5% or 8% respectively. It must be noted that most data points in the knocked down environments are outside the support of the observational training data such that $\hat{f}(\mathbf{X})$ can also be a poor approximation for causal effects; see also the discussion regarding out-of-support interventions in Section 2.4. Hence, this analysis of the regression performance in other environments, although in line with our other results, shall be viewed with some caution. The analysis for this target variable corresponds to the last row-box in Table 1.

Of course, other genes could be viewed as target Y. When estimating a Markov blanket as described above for different variables, the interventional environments often indicate the existence of cyclicity between the target and all its potential causes. Then, our method is of little help as the different predictors cannot be grouped into different classes. In Table 1, we summarize the results for all possible targets with multiple predictors where at least one predictor appears to be neither a descendant of the target nor in a cyclic relation using a threshold of 0.01 for the Mann-Whitney U test. In 4 out of 6 cases, the ranking implied by our method in terms of number of splits where a predictor is selected by FOCI is in agreement with the ranking implied by the Mann-Whitney U test, and $\hat{W}$ using $\tilde{\alpha} = 0.01$ is exactly as implied by the interventional data. Of the remaining two cases, the method is

Target Y	Predictor X_j	Mann-Whitney U	Splits	Proportion test	Relative bias
X_5	X_2	2.3e-18	14	1.2e-02	1.3e-01
	X_3	4.9e-31	26	–	2.1e-02
	X_4	1.2e-69	19	1.1e-01	3e-02
	X_{12}	3.5e-01	28	–	4.3e-02
X_6	X_{11}	7.7e-01	17	2.3e-05	1e-01
	X_{24}	2.4e-09	38	–	1.5e-01
X_7	X_8	3.3e-02	28	–	1e-01
	X_9	1.4e-16	10	2e-04	8.3e-02
	X_{14}	1.2e-79	30	–	9.7e-02
	X_{22}	1.6e-35	11	4.6e-04	4.1e-02
X_9	X_7	5.4e-01	14	2.2e-03	1.8e-02
	X_{11}	1.2e-14	30	–	1.8e-02
	X_{22}	2.3e-06	29	–	2.4e-02
X_{15}	X_{11}	2.3e-02	12	4.9e-07	1.1e-01
	X_{16}	1.6e-06	37	–	1.2e-01
X_{16}	X_{10}	1e-01	22	7.7e-05	4.7e-02
	X_{12}	4.7e-01	19	6.3e-06	7.7e-02
	X_{15}	3.8e-05	41	–	1.2e-01

Table 1: Application to the K562 data set with varying targets and predictor sets. The third column is the p-value of the Mann-Whitney U test comparing the observational distribution of the predictor to its distribution when knocking down the target. The fourth and fifth column report the output of Algorithm 3, i.e., the number of splits where FOCI selects this predictor, n_j, and the p-value of the proportion test if $n_j < \bar{n}$ (the significant findings with small p-value correspond to the variables which are causally well-specified; no p-value indicates that $n_j \geq \bar{n}$ and the variable is not causally well-specified). The last column reports the relative bias $RB^{X_j \to Y}$ (7) when using the model fit on observational data to predict the target in the data set where the predictor is knocked down.

once conservative $\hat{W} = \emptyset$ (for $Y = X_5$) and once the interventional data suggest that there are false positives in $\hat{W}$ (for $Y = X_7$). $Y = X_{16}$ is the case discussed in more detail above.

6. Location-Scale Noise Models

A simple extension of model (2), that has recently gained some attention, is the heteroskedastic noise model also referred to as the location-scale noise model (LSNM). There, the independent, additive noise is scaled by some nonnegative function $g_{\mathbf{X}Y}(\mathbf{X}_{\mathrm{PA}(Y)})$ inducing heteroskedasticity. This is the leading causal model in, e.g., Xu et al. (2022); Strobl and Lasko (2023); Immer et al. (2023), where the latter two provide identifiability guarantees. In analogy to (2), we call the LSNM causally well-specified if

$$Y \leftarrow f_{\mathbf{X}Y}(\mathbf{X}_{\mathrm{PA}(Y)}) + g_{\mathbf{X}Y}(\mathbf{X}_{\mathrm{PA}(Y)}) f_{\mathbf{H}Y}(\mathbf{H}_{\mathrm{PA}(Y)}), \quad \text{where} \quad \mathbf{H}_{\mathrm{PA}(Y)} \perp \mathbf{X}. \qquad (8)$$

We choose the parametrization $\mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}\big)\big] = 0$ and $\mathbb{E}\Big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}\big)^2\Big] = 1$ such that $f_{\mathbf{X}Y}(\cdot)$ and $g^2_{\mathbf{X}Y}(\cdot)$ denote the conditional mean and variance. As before, the independence condition implies

$$Y|\mathbf{X} = \mathbf{x} \quad \overset{d}{=} \quad Y|\mathrm{do}(\mathbf{X} \leftarrow \mathbf{x}).$$

With the others, one can separate the independent noise term such that one can understand the counterfactual of changing the predictors.

$$P_Y^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\mathrm{do}(\mathbf{X} \leftarrow \mathbf{x}')} = \delta_{y'} \quad \text{where}$$
$$y' = (y - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}])\sqrt{\mathrm{Var}(Y|\mathbf{X} = \mathbf{x}')/\mathrm{Var}(Y|\mathbf{X} = \mathbf{x})} + \mathbb{E}\big[Y|\mathbf{X} = \mathbf{x}'\big].$$

To check (8), we have the natural proxy

$$H_0: \quad \mathcal{E} \perp \mathbf{X}, \quad \text{where} \quad \mathcal{E} = \frac{Y - \mathbb{E}[Y|\mathbf{X}]}{\sqrt{\mathrm{Var}(Y|\mathbf{X})}} \tag{9}$$

since under (8) we have that $\mathcal{E} = f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}\big)$ and hence H_0 in (9) holds.

In case of model misspecification, we can consider the per-covariate causal well-specification. Condition (A1) remains the same for the LSNM, (A2) can be replaced by a weaker version for this more flexible causal model:

(A2*) $Y \leftarrow f_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big) + g_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big) f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)$, i.e., with addition and multiplication of measured functions, one can separate a term that does not include $\mathbf{X}_U$.

Again, these assumptions imply a counterfactual statement and a testable proxy.

Theorem 5. Assume the model (1) with (B1). Let $\mathbf{X}_U$ be a set of covariates fulfilling (A1) and (A2*), then

$$P_Y^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\mathrm{do}\big(\mathbf{X}_U \leftarrow \mathbf{x}'_U, \mathbf{X}_{-U} \leftarrow \mathbf{X}_{-U}\big)} = \delta_{y'} \quad \text{where}$$
$$y' = (y - \mathbb{E}[Y|\mathbf{X} = \mathbf{x}])\sqrt{\mathrm{Var}\big(Y|\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{X}_{-U}\big)/\mathrm{Var}(Y|\mathbf{X} = \mathbf{x})} +$$
$$\mathbb{E}\big[Y|\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{X}_{-U}\big]$$

for $(\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U})$ in the support of the observational distribution. Further, $H_{0,U}$ holds, where

$$H_{0,U}: \quad \mathcal{E} \perp \mathbf{X}_U|\mathbf{X}_{-U}, \quad \text{with} \quad \mathcal{E} = \frac{Y - \mathbb{E}[Y|\mathbf{X}]}{\sqrt{\mathrm{Var}(Y|\mathbf{X})}}. \tag{10}$$

By constructing a counterfactual such that the regression residual $\mathcal{E}$ remains unchanged, the effect on Y can be assessed in terms of the conditional mean and the conditional variance. As in Section 2.4 one could alternatively use do-statements for $(\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U})$ outside the support of the observational distribution.

6.1 Asymptotic Results

To fit location-scale noise models, a simple approach is to estimate both $\mathbb{E}[Y|\mathbf{X}]$ and $\mathbb{E}[Y^2|\mathbf{X}]$. If both these quantities are known, one can recover $\mathcal{E}$.

We consider variations of Algorithms 1 and 2 where we get estimates $\hat{f}_1(\mathbf{X})$ for $f_1(\mathbf{X}) \coloneqq \mathbb{E}[Y|\mathbf{X}]$ and $\hat{f}_2(\mathbf{X})$ for $f_2(\mathbf{X}) \coloneqq \mathbb{E}[Y^2|\mathbf{X}]$ using certain regressors on the data $(\mathbf{x}, \mathbf{y})$; see the notation in Section 3. Then, we estimate the residuals

$$\epsilon_i = \frac{y_i - f_1(\mathbf{x}_i)}{\sqrt{f_2(\mathbf{x}_i) - f_1^2(\mathbf{x}_i)}} \quad \text{by} \quad \hat{\epsilon}_i = \frac{y_i - \hat{f}_1(\mathbf{x}_i)}{\sqrt{\hat{f}_2(\mathbf{x}_i) - \hat{f}_1^2(\mathbf{x}_i)}}.$$

Especially for low sample sizes, it can happen that $\hat{f}_2(\mathbf{x}_i) \leq \hat{f}_1^2(\mathbf{x}_i)$ for some i. To make the method operational in such cases, we suggest defining $\hat{\epsilon}_i$ by a large quantity in absolute value with the same sign as $y_i - \hat{f}_1(\mathbf{x}_i)$. For our asymptotic results, it could even be replaced by arbitrary values. To establish guarantees for FOCI, we make the following assumptions

(B6) $\left|f_1(\mathbf{x}_i) - \hat{f}_1(\mathbf{x}_i)\right| = \mathcal{O}_p(1)$.

(B7) $\left|f_2(\mathbf{x}_i) - \hat{f}_2(\mathbf{x}_i)\right| = \mathcal{O}_p(1)$.

(B8) $\mathbb{P}\big(f_2(\mathbf{x}_i) - f_1^2(\mathbf{x}_i) > 0\big) = 1$.

In Assumptions (B6) and (B7) the probability is over both, the function estimates and the new data point. Assumption (B8) implies that Y is almost surely not deterministic in $\mathbf{X}$.

Theorem 6. Suppose that the regularity assumptions (A1') and (A2') (Azadkia and Chatterjee, 2021) for the data $(g(\mathcal{E}), \mathbf{X})$ hold as well as conditions (B2) and (B4) - (B8). Let $\hat{S}$ be the output of Algorithm 2 modified to normalize the residuals for the heteroscedastic noise model. There are positive real numbers L_1, L_2 and L_3 that do not depend on the sample size such that

$$\mathbb{P}\Big(\hat{S} \supseteq \{1, \ldots, p\} \setminus W\Big) \geq 1 - L_1 p^{L_2} \exp(-L_3 n).$$

If instead $\hat{S}$ is the output of Algorithm 1 adjusted to normalize the residuals, it holds

$$\lim_{n \to \infty} \mathbb{P}\Big(\hat{S} \supseteq \{1, \ldots, p\} \setminus W\Big) = 1.$$

The key step to adapt the results to the heteroskedastic case is seeing that (B6) - (B8) imply

$$|\hat{\epsilon}_i - \epsilon_i| = \mathcal{O}_p(1).$$

Then, all the results from the homoskedastic case carry over. Any other regression algorithm tailor-made for location-scale noise models could be applied as well if it ensures this condition.

Although we receive similar asymptotic guarantees for location-scale noise models under rather weak assumptions, they are harder to deal with for finite samples. As all conditional

dependence between Y and any X_j that is due to location or scale is regressed out, the residing dependence can be very weak. Hence, the population Conditional Dependence Coefficient (Azadkia and Chatterjee, 2021) is low requiring an even larger sample size. Also, the absolute value transform appears to be less appropriate after regressing away the scale information. Hence, we apply no transform in the simulation example.

6.2 Simulation Example

We consider a simple example with two observed predictors and one hidden confounder as shown in Figure 5. We let

$$Y \leftarrow g_{X_2Y}(X_2)H$$

such that (A2*) holds for X_2. The causal effect is sinusoidal from H to X_1, linear from X_1 to X_2, and there is an additive Gaussian error term on each.

For each sample size from 10^2 to 10^5, we run 200 repetitions of the same data generating mechanism. We fit both moments with `xgboost` and use the identity function for $g(\cdot)$. Otherwise, we proceed as in Section 4.

In Figure 6, we show the same performance metrics as in Figure 4. We see that our method can handle this toy example quite well. For 10^5 samples, the performance with $\tilde{\alpha} = 0.01$ is almost perfect, i.e., 196 times the output is $\hat{W} = \{2\}$ and 4 times $\hat{W} = \emptyset$. There are no false positives in $\hat{W}$.

The global test works well already for 10^2 samples. After aggregation over the different splits, H_0 in (9) is rejected in every simulation run at $\alpha = 0.05$. This can be facilitated by the fact that the fits are not good for this sample size such that there is more dependence on $\mathbf{x}$ for $\hat{\epsilon}$ than for the true ϵ.

Finally, we compare how well the ordering of $\hat{\boldsymbol{\epsilon}}$ matches that of $\boldsymbol{\epsilon}$. For each run, we calculate the average misposition defined as

$$AMP = \frac{1}{n^2} \sum_{i=1}^{n} \left| \sum_{l=1}^{n} \mathbb{1}_{\{\epsilon_l < \epsilon_i\}} - \mathbb{1}_{\{\hat{\epsilon}_l < \hat{\epsilon}_i\}} \right|. \tag{11}$$

We show the according box plots on the right side of Figure 6. As desired, this quantity approaches 0 for increasing sample size. For simplicity, we calculate this quantity only on a single split per simulation run.
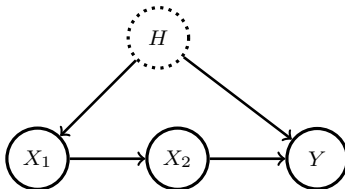

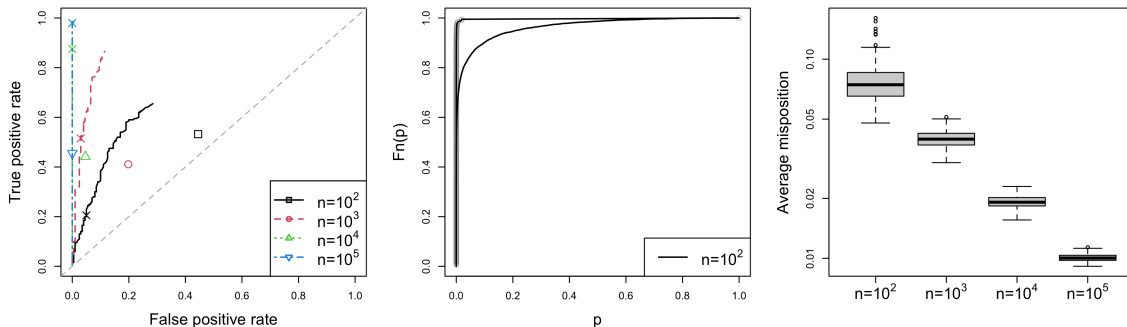Figure 5: DAG representing the SCM in the simulation fitting LSNM.

Figure 6: The results are based on 200 simulation runs. On the left: False positive rate versus true positive rate obtained with Algorithm 3 adjusted to LSNM for varying $\tilde{\alpha}$ and $\alpha = 0.05$. The crosses correspond to $\tilde{\alpha} = 0.01$. The other symbols describe the performance of Algorithm 2 adjusted to LSNM. In the middle: empirical cumulative distribution function of the p-values obtained with HSIC. We compare the raw p-values from each split (lines only) to the cumulated p-value per simulation run (lines with dots). On the right: average misposition (11) of the estimated residuals with respect to the true residuals.

7. Conclusion

In this paper, we introduce the notion of causal well-specification for additive noise models or their extension to heteroskedastic errors. Our viewpoint of local, i.e., for a subset of the covariates, causal well-specification, for which conditional independence between predictor and residual can serve as a proxy, provides a new option instead of rejecting entire models.

We present an algorithm to estimate our quantities of interest from finite data and provide some asymptotic guarantees. We demonstrate its application in simulation setups. This reveals some difficulties but also shows how considering multiple data splits can help even in hard cases.

Finally, we also apply our methodology and algorithm to regression problems extracted from a large-scale genomic data set. While in many cases, causal well-specification appears to be not even approximately fulfilled, we find multiple examples where our estimate of well-specification is in line with an approximate validation from various gene knockdown perturbations.

We would like to emphasize that our formulation and analysis of the information provided by conditional independence, which we present in Section 2, can also be applied as stand-alone and other machine learning methods for regression and conditional dependency assessment can be used.

Code scripts to reproduce the results presented in this paper are available here
https://github.com/cschultheiss/nl_GOF.

Acknowledgments

Appendix A. Proofs

We provide here the proofs for the results stated in the main text.

A.1 Proof of Theorem 1

Recall
$$Y := \mathbb{E}[Y|\mathbf{X}] + \mathcal{E}.$$

Due to (A2), we have

$$\mathbb{E}[Y|\mathbf{X}] = f_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big) + \mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big] \quad \text{such that}$$
$$\mathcal{E} = f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big) - \mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big].$$

Using (A1), $\mathbf{X}_U \perp \mathbf{H}_{\mathrm{PA}(Y)}|\mathbf{X}_{-U}$, and trivially, $\mathbf{X}_U \perp \mathbf{X}_{-U}|\mathbf{X}_{-U}$. It follows

$$\mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big] = \mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}_{-U}\big] \perp \mathbf{X}_U|\mathbf{X}_{-U} \quad \text{and}$$
$$f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus j}\big) \perp \mathbf{X}_U|\mathbf{X}_{-U} \quad \text{such that} \quad \mathcal{E} \perp \mathbf{X}_U|\mathbf{X}_{-U}.$$

Consider the counterfactual intervention. As $\mathbf{X}_{-U}$ remains unchanged, the second summand in (A2) could only change if $\mathbf{H}_{\mathrm{PA}(Y)}$ changes. This could happen through some directed path from $\mathbf{X}_U$ to $\mathbf{H}_{\mathrm{PA}(Y)}$ that is not blocked by $\mathbf{X}_{-U}$. By (A1), if such an effect from $\mathbf{X}_U$ to $\mathbf{H}_{\mathrm{PA}(Y)}$ exists, it is constant for almost all $\mathbf{x}_U$. With (B1), we can extend this argument to all attainable $\mathbf{x}_U$. Hence, changing $\mathbf{X}_U$ from $\mathbf{x}_U$ to $\mathbf{x}'_U$ while keeping $\mathbf{X}_{-U}$ fixed, cannot affect $\mathbf{H}_{\mathrm{PA}(Y)}$ such that the second summand remains constant. For the first summand, we can directly plug in the counterfactual values of $\mathbf{X}$.

In the conditional expectation given above, only the first summand can change as the second is a function of only $\mathbf{X}_{-U}$. As the altered summand is the same for both Y and $\mathbb{E}[Y|\mathbf{X}]$, the new value y' must exactly represent this change in conditional mean. ∎

A.2 Proof of Theorem 2

Consider first the $\subseteq$-statement. This means that $H_{0,j}$ in (6) must hold $\forall j \in W$. Let $S = \{1, \ldots, p\} \setminus W$. Then, we want that

$$\mathcal{E} \perp \mathbf{X}_W|\mathbf{X}_S \implies \mathcal{E} \perp X_j|\mathbf{X}_{-j}.$$

This can be rewritten as

$$\mathcal{E} \perp \mathbf{X}_{W\setminus j}, X_j|\mathbf{X}_S \implies \mathcal{E} \perp X_j|\mathbf{X}_S, \mathbf{X}_{W\setminus j}.$$

This is the weak union property in Chapter 3 of Pearl (1988) and hence holds for any random variables.

For $W = \tilde{W}$, we additionally need that $H_{0,j}$ cannot hold for any $j \in S$. By minimality of S

$$\mathcal{E} \not\perp X_j, \mathbf{X}_W|\mathbf{X}_{S\setminus j}.$$

Then, the intersection property implies

$$\mathcal{E} \not\perp \mathbf{X}_W|X_j, \mathbf{X}_{S\setminus j} \quad \text{or} \quad \mathcal{E} \not\perp X_j|\mathbf{X}_W, \mathbf{X}_{S\setminus j}.$$

The first cannot hold by the definition of W, so the second must hold. This means that $H_{0,j}$ is not fulfilled, and $W = \tilde{W}$ is guaranteed. As $\tilde{W}$ is unique by construction, W is then unique as well. ∎

A.3 Definitions from FOCI

These definitions are taken from (Azadkia and Chatterjee, 2021), adapted in parts to fit our notation. Let μ be the law of $\mathcal{E}$. We have the following population quantities

$$Q(\mathcal{E}, \mathbf{X}_U) = \int \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t | \mathbf{X}_U)) d\mu(t) \geq 0$$

$$S(\mathcal{E}) = \int \mathrm{Var}(\mathbb{1}_{\mathcal{E} \geq t}) d\mu(t) \geq 0$$

$$T(\mathcal{E}, \mathbf{X}_U) = Q(\mathcal{E}, \mathbf{X}_U)/S(\mathcal{E}) \in [0, 1]$$

$$Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) = \int \mathbb{E}[\mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t | \mathbf{X}_U, \mathbf{X}_S) | \mathbf{X}_S)] d\mu(t) \geq 0$$

$$S(\mathcal{E}, \mathbf{X}_S) = \int \mathbb{E}[\mathrm{Var}(\mathbb{1}_{\mathcal{E} \geq t} | \mathbf{X}_S)] d\mu(t) \geq 0$$

$$T(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) = Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S)/S(\mathcal{E}, \mathbf{X}_S) \in [0, 1].$$

For data estimates, define first

$$R_i = \sum_{l=1}^n \mathbb{1}_{\epsilon_l \leq \epsilon_i}, \quad L_i = \sum_{l=1}^n \mathbb{1}_{\epsilon_l \geq \epsilon_i}$$

and $M(i)$ the nearest neighbour of i with respect to $\mathbf{X}_U$ with a random tie-breaking rule. Then, we have the data estimates

$$Q_n(\boldsymbol{\epsilon}, \mathbf{x}_U) = \frac{1}{n^2} \sum_{i=1}^n \min\{R_i, R_{M(i)}\} - \frac{L_i^2}{n}$$

$$S_n(\boldsymbol{\epsilon}) = \frac{1}{n^3} \sum_{i=1}^n L_i(n - L_i)$$

$$T_n(\boldsymbol{\epsilon}, \mathbf{x}_U) = Q_n(\boldsymbol{\epsilon}, \mathbf{x}_U)/S_n(\boldsymbol{\epsilon})$$

$$S_n(\boldsymbol{\epsilon}, \mathbf{x}_U) = \frac{1}{n^2} \sum_{i=1}^n R_i - \min\{R_i, R_{M(i)}\}.$$

A.4 Proof of Proposition 1

We have $T(\cdot, \mathbf{X}_S) = Q(\cdot, \mathbf{X}_S)/S(\cdot)$. As argued in Azadkia and Chatterjee (2021), the denominator for unconditional independence tests is simply $1/6$ for continuous random variables. If $\mathcal{E}$ is conditionally continuously distributed, the same holds for its marginal distribution and thus also for the distribution of $|\mathcal{E}|$. Hence, it suffices to consider $Q(\cdot, \mathbf{X}_S)$ and the statement for $T(\cdot, \mathbf{X}_S)$ follows directly. Let μ and ν be the law of $\mathcal{E}$ and $|\mathcal{E}|$. Due

to symmetry, it holds $d\nu(t) = 2d\mu(t) \ \forall t \geq 0$.

$$Q(|\mathcal{E}|, \mathbf{X}_S) = \int_0^\infty \text{Var}(\mathbb{P}(|\mathcal{E}| \geq t|\mathbf{X}_S))d\nu(t) = \int_0^\infty \text{Var}(2\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))2d\mu(t)$$

$$= 8\int_0^\infty \text{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t),$$

$$\int_{-\infty}^0 \text{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t) = \int_{-\infty}^0 \text{Var}(\mathbb{P}(\mathcal{E} \leq -t|\mathbf{X}_S))d\mu(t)$$

$$= \int_{-\infty}^0 \text{Var}(1 - \mathbb{P}(\mathcal{E} \geq -t|\mathbf{X}_S))d\mu(t) = \int_{-\infty}^0 \text{Var}(\mathbb{P}(\mathcal{E} \geq -t|\mathbf{X}_S))d\mu(t)$$

$$\stackrel{t' \Leftarrow -t}{=} \int_0^\infty \text{Var}(\mathbb{P}(\mathcal{E} \geq t'|\mathbf{X}_S))d\mu(t') \text{ such that}$$

$$Q(\mathcal{E}, \mathbf{X}_S) = \int_{-\infty}^\infty \text{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t) = 2\int_0^\infty \text{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t).$$

The first line uses symmetry, and the second chain of equalities uses symmetry as well as continuity to allow for a weak inequality in the complementary probability. Comparing the quantity on the first line to that on the last line we see that the ratio between the numerator terms is 4. ∎

A.5 Proof of Theorem 3

We build up the proof by some supporting Lemmata.

Lemma 1. Assume (B3) and (B4). Then, $\forall i \neq l$

$$\lim_{n \to \infty} \mathbb{P}([g(\epsilon_i) > g(\epsilon_l) \cap g(\hat{\epsilon}_i) \leq g(\hat{\epsilon}_l)] \cup [g(\epsilon_i) < g(\epsilon_l) \cap g(\hat{\epsilon}_i) \geq g(\hat{\epsilon}_l)] \cup [g(\epsilon_i) = g(\epsilon_l)]) = 0,$$

i.e., the probability that the estimates imply a different ordering between i and l approaches 0.

Define $Q_n(\cdot)$ and $S_n(\cdot)$ as in Section 9 of Azadkia and Chatterjee (2021).

Lemma 2. Assume the conditions of Lemma 1. Let U be any non-empty subset of $\{1, \ldots, p\}$. Then,

$$\lim_{n \to \infty} \mathbb{E}[|Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)|] = 0, \ \lim_{n \to \infty} \mathbb{E}[|S_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - S_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)|] = 0,$$

$$\lim_{n \to \infty} \mathbb{E}[|S_n(g(\boldsymbol{\epsilon})) - S_n(g(\hat{\boldsymbol{\epsilon}}))|] = 0.$$

As in the sample splitting case $(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)$ are i.i.d. copies, one can apply Lemma 11.9 in Azadkia and Chatterjee (2021) to those. This yields

$$\mathbb{P}(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - \mathbb{E}[Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)]| \geq t) \leq K_1 \exp(-K_2 n t^2), \tag{12}$$

for some positive K_1, K_2. Therefore, we can draw similar conclusions as in their Lemma 14.2.

Lemma 3. Let U be a subset of size u. Assume conditions (A1), which defines β, and (A2) from Azadkia and Chatterjee (2021) for the data $(g(\mathcal{E}), \mathbf{X}_U)$ as well as conditions (B3) - (B4). Then, there exist positive K_1, K_2, and K_3 that do not depend on the sample size such that in the sample splitting case

$$\mathbb{P}\Big(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)| \geq K_1 \max\Big\{D^{1/3}(n), n^{-\min\{-1/u, -1/2\}} \log(n)^{u+\beta+1}\Big\} + t\Big) \leq$$
$$K_2 \exp\big(-K_3 n t^2\big).$$

Under (B2) and (B5) any set U that is not a (weak) superset of $\{1, \ldots, p\} \setminus W$ cannot be sufficient for $g(\mathcal{E})$. Thus, it suffices to bound the probability of $\hat{S}$ not being sufficient, and then Theorem 3 follows. This corresponds to Theorem 6.1 in Azadkia and Chatterjee (2021). The only part of its proof that needs adaptation is Lemma 16.3. To proof an according result based on our Lemma 3, we require

$$L_1 \max\Big\{D(n), n^{-\min\{-1/K, -1/2\}} \log(n)^{K+\beta+1}\Big\} \leq \frac{\delta}{16}.$$

Here, we use their definition of δ, i.e., δ is the largest number such that for any insufficient subset $U \not\supseteq (\{1, \ldots, p\} \setminus W)$, there exists $j \notin U$ that fulfils $Q(g(\mathcal{E}), \mathbf{X}_{U \cup j}) \geq Q(g(\mathcal{E}), \mathbf{X}_U) + \delta$. K is the integer part of $1/\delta + 2$. As we consider fixed data generating mechanisms, $\delta > 0$ holds by construction. Hence, we do not mention it in the theorems explicitly. This inequality might require a larger sample size than in Azadkia and Chatterjee (2021) and larger L_6 accordingly. Apart from that, the proof follows from the same principles. ∎

A.5.1 Proof of Lemma 1

The properties of $g(\cdot)$ imply that $g(\mathcal{E})$ is a continuous random variable as well such that the probability of the last event has probability 0 regardless of the sample size. As i and l are interchangeable, the first two events have the same probability and it suffices to analyse one. Let $\eta > 0$ be arbitrary.

$$\mathbb{P}(g(\epsilon_i) > g(\epsilon_l) \cap g(\hat{\epsilon}_i) \leq g(\hat{\epsilon}_l)) =$$
$$\mathbb{P}(g(\epsilon_i) > g(\epsilon_l) \cap g(\hat{\epsilon}_i) \leq g(\hat{\epsilon}_l) \cap g(\epsilon_i) - g(\epsilon_l) \leq \eta) +$$
$$\mathbb{P}(g(\epsilon_i) > g(\epsilon_l) \cap g(\hat{\epsilon}_i) \leq g(\hat{\epsilon}_l) \cap g(\epsilon_i) - g(\epsilon_l) > \eta) \leq$$
$$\mathbb{P}(|g(\epsilon_i) - g(\epsilon_l)| \leq \eta) + \mathbb{P}(|g(\hat{\epsilon}_i) - g(\epsilon_i)| + |g(\hat{\epsilon}_l) - g(\epsilon_l)| > \eta) \leq$$
$$\mathbb{P}(|g(\epsilon_i) - g(\epsilon_l)| \leq \eta) + \mathbb{P}(\max\{|g(\hat{\epsilon}_i) - g(\epsilon_i)|, |g(\hat{\epsilon}_l) - g(\epsilon_l)|\} > \eta/2) \leq$$
$$\mathbb{P}(|g(\epsilon_i) - g(\epsilon_l)| \leq \eta) + 2\mathbb{P}(|g(\hat{\epsilon}_i) - g(\epsilon_i)| > \eta/2) \leq$$
$$\mathbb{P}(|g(\epsilon_i) - g(\epsilon_l)| \leq \eta) + 2\mathbb{P}(|\hat{\epsilon}_i - \epsilon_i| > \eta/2l).$$

Let now η depend on n. For $\eta \to 0$ the first term vanishes. If it approaches 0 slowly enough, the second term vanishes as well assuming the regression is suitable. Thus, one can choose η such that both terms vanish. Since the inequality holds for arbitrary η, the probability goes to 0, i.e.,

$$\mathbb{P}([g(\epsilon_l) \leq g(\epsilon_i) \cap g(\hat{\epsilon}_l) > g(\hat{\epsilon}_i)] \cup [g(\epsilon_l) \geq g(\epsilon_i) \cap g(\hat{\epsilon}_l) < g(\hat{\epsilon}_i)]) = \mathcal{O}(1). ∎$$

A.5.2 Proof of Lemma 2

Let $R_i = \sum g(\epsilon_l) \leq g(\epsilon_i)$, $L_i = \sum g(\epsilon_l) \geq g(\epsilon_i)$, and $\hat{R}_i$, $\hat{L}_i$ the according quantities estimated with $\hat{\epsilon}$. Note that index $M(i)$, i.e., the nearest neighbour of i with respect to $\mathbf{x}_U$, only depends on observed quantities. Hence, it is the same for the estimated quantity $\hat{R}_{M(i)}$.

$$|Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)| = \left| \frac{1}{n^2} \sum_{i=1}^{n} \min\{R_i, R_{M(i)}\} - \min\{\hat{R}_i, \hat{R}_{M(i)}\} + \frac{\hat{L}_i^2 - L_i^2}{n} \right| \leq$$

$$\left| \frac{1}{n^2} \sum_{i=1}^{n} \min\{R_i, R_{M(i)}\} - \min\{\hat{R}_i, \hat{R}_{M(i)}\} \right| + \left| \frac{1}{n^3} \sum_{i=1}^{n} \hat{L}_i^2 - L_i^2 \right|.$$

$$|S_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - S_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)| = \left| \frac{1}{n^2} \sum_{i=1}^{n} R_i - \hat{R}_i + \min\{\hat{R}_i, \hat{R}_{M(i)}\} - \min\{R_i, R_{M(i)}\} \right| \leq$$

$$\left| \frac{1}{n^2} \sum_{i=1}^{n} R_i - \hat{R}_i \right| + \left| \frac{1}{n^2} \sum_{i=1}^{n} \min\{\hat{R}_i, \hat{R}_{M(i)}\} - \min\{R_i, R_{M(i)}\} \right|.$$

$$|S_n(g(\boldsymbol{\epsilon})) - S_n(g(\hat{\boldsymbol{\epsilon}}))| = \left| \frac{1}{n^3} \sum_{i=1}^{n} n\left(L_i - \hat{L}_i\right) + \hat{L}_i^2 - L_i^2 \right| \leq$$

$$\left| \frac{1}{n^2} \sum_{i=1}^{n} L_i - \hat{L}_i \right| + \left| \frac{1}{n^3} \sum_{i=1}^{n} \hat{L}_i^2 - L_i^2 \right|.$$

Thus, there are four different terms to be controlled. If both $\boldsymbol{\epsilon}$ and $\hat{\boldsymbol{\epsilon}}$ have n distinct values, all the terms that do not depend on the nearest neighbouring property amongst $\mathbf{x}_U$ are trivially 0 for all sample sizes. However, we can prove convergence without this assumption.

$$\mathbb{E}\left[\left| \frac{1}{n^2} \sum_{i=1}^{n} R_i - \hat{R}_i \right| \right] \leq \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \left| R_i - \hat{R}_i \right| \right] =$$

$$\mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \left| \sum_{l=1}^{n} \mathbb{1}_{\{g(\epsilon_l) \leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l) \leq g(\hat{\epsilon}_i)\}} \right| \right] =$$

$$\mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \left| \sum_{l \neq i} \mathbb{1}_{\{g(\epsilon_l) \leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l) \leq g(\hat{\epsilon}_i)\}} \right| \right] \leq$$

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{l \neq i} \mathbb{E}\left[\left| \mathbb{1}_{\{g(\epsilon_l) \leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l) \leq g(\hat{\epsilon}_i)\}} \right| \right] =$$

$$\frac{n^2 - n}{n^2} \mathbb{E}\left[\left| \mathbb{1}_{\{g(\epsilon_l) \leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l) \leq g(\hat{\epsilon}_i)\}} \right| \right] =$$

$$\frac{n^2 - n}{n^2} \mathbb{P}([g(\epsilon_l) \leq g(\epsilon_i) \cap g(\hat{\epsilon}_l) > g(\hat{\epsilon}_i)] \cup [g(\epsilon_l) \geq g(\epsilon_i) \cap g(\hat{\epsilon}_l) < g(\hat{\epsilon}_i)]) \xrightarrow{n \to \infty} 0$$

by Lemma 1. In the last two expressions, $l \neq i$ is assumed. The argument for the term with $L_i - \hat{L}_i$ is identical.

$$\mathbb{E}\left[\left|\frac{1}{n^2}\sum_{i=1}^{n}\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} - \min\left\{R_i, R_{M(i)}\right\}\right|\right] \leq$$

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left|\hat{R}_i - R_i\right| + \left|\hat{R}_{M(i)} - R_{M(i)}\right|\right] =$$

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left|\hat{R}_i - R_i\right| + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l:\ M(l)=i}\left|\hat{R}_i - R_i\right|\right] =$$

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left|\hat{R}_i - R_i\right|\left(1 + \sum_{l\neq i}\mathbb{1}_{M(l)=i}\right)\right] =$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left|\hat{R}_i - R_i\right|\left(1 + \sum_{l\neq i}\mathbb{1}_{M(l)=i}\right)\right] =$$

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left|\hat{R}_i - R_i\right|\mathbb{E}\left[1 + \sum_{l\neq i}\mathbb{1}_{M(l)=i}|\hat{R}_i, R_i\right]\right] \leq \frac{2 + C(p)}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left|\hat{R}_i - R_i\right|\right] \overset{n\to\infty}{\to} 0.$$

By Lemma 11.4 in Azadkia and Chatterjee (2021), there is a dimension-dependent constant such that no point can be the nearest neighbour of more than $C(p)$ distinct points in $\mathbb{R}^p$. If there are l such that $\mathbf{x}_{l,U} = \mathbf{x}_{i,U}$, $M(l)$ is chosen uniformly at random from this set, and in expectation there is one l such that $M(l) = i$. As this uniform draw is independent of R_i and $\hat{R}_i$, the upper bound also applies to the conditional expectation and we can pull it out.

$$\mathbb{E}\left[\left|\frac{1}{n^3}\sum_{i=1}^{n}L_i^2 - \hat{L}_i^2\right|\right] = \mathbb{E}\left[\left|\frac{1}{n^3}\sum_{i=1}^{n}\left(L_i - \hat{L}_i\right)\left(L_i + \hat{L}_i\right)\right|\right] \leq \mathbb{E}\left[\left|\frac{2}{n^2}\sum_{i=1}^{n}L_i - \hat{L}_i\right|\right] \overset{n\to\infty}{\to} 0.$$

Thus, every term is under control which concludes the proof. As all the terms are at most of the same order as the probability in Lemma 1, the bound on the convergence rate follows directly. ∎

A.5.3 Proof of Lemma 3

By Lemma 2, there exists a rate, say, $D(n) = \mathcal{O}(1)$ such that

$$|\mathbb{E}[Q_n(g(\hat{\epsilon}), \mathbf{x}_U)] - \mathbb{E}[Q_n(g(\epsilon), \mathbf{x}_U)]| = \mathcal{O}(D(n)).$$

Then,

$$|Q_n(g(\hat{\epsilon}), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)| \leq$$
$$|Q_n(g(\hat{\epsilon}), \mathbf{x}_U) - \mathbb{E}[Q_n(g(\hat{\epsilon}), \mathbf{x}_U)]| + |\mathbb{E}[Q_n(g(\hat{\epsilon}), \mathbf{x}_U)] - \mathbb{E}[Q_n(g(\epsilon), \mathbf{x}_U)]| +$$
$$|\mathbb{E}[Q_n(g(\epsilon), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)]| \leq$$
$$|Q_n(g(\hat{\epsilon}), \mathbf{x}_U) - \mathbb{E}[Q_n(g(\hat{\epsilon}), \mathbf{x}_U)]| + K_1 \max\left\{D(n), n^{-\min\{-1/u, -1/2\}}\log(n)^{u+\beta+1}\right\},$$

using the rate derived in Lemma 14.2 of (Azadkia and Chatterjee, 2021). Therefore, with (12),

$$\mathbb{P}\Big(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)| \geq K_1 \max\Big\{D(n), n^{-\min\{-1/u, -1/2\}} \log(n)^{u+\beta+1}\Big\} + t\Big) \leq$$
$$\mathbb{P}(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - \mathbb{E}[Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)]| \geq t) \leq K_2 \exp\big(-K_3 n t^2\big). \ \blacksquare$$

A.6 Proof of Theorem 4

Again, we only have to bound the probability of $\hat{S}$ not being sufficient.

Using Lemma 2 and the Markov inequality, we see

$$\mathbb{P}(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U)| \geq t) \leq \frac{K_1 D(n)}{t}.$$

Hence, we get

$$\mathbb{P}\Big(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)| \geq K_1 n^{-\min\{-1/u, -1/2\}} \log(n)^{u+\beta+1} + t\Big) \leq$$
$$\mathbb{P}\Big(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U)| + |Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)| \geq$$
$$K_1 n^{-\min\{-1/u, -1/2\}} \log(n)^{u+\beta+1} + t\Big) \leq$$
$$\mathbb{P}\Big(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U)| \geq \frac{t}{2}\Big) +$$
$$\mathbb{P}\Big(|Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - Q(g(\mathcal{E}), \mathbf{X}_U)| \geq K_1 n^{-\min\{-1/u, -1/2\}} \log(n)^{u+\beta+1} + \frac{t}{2}\Big) \leq$$
$$\frac{K_2(n)}{t} + K_3 \exp\big(-K_4 n t^2\big) \leq K_5 \max\Big\{\frac{D(n)}{t}, \exp\big(-K_4 n t^2\big)\Big\},$$

where we used Lemma 14.2 in (Azadkia and Chatterjee, 2021) in the second to last inequality. Finally, we can follow the proof idea of Lemma 16.3 in (Azadkia and Chatterjee, 2021) with the given probability bound showing that the probability of $\hat{S}$ being insufficient goes to 0. $\blacksquare$

A.7 Proof of Proposition 2

For the least squares parameter, we have

$$\hat{\beta} = \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} = \beta + \frac{\mathbf{x}^\top \boldsymbol{\epsilon}}{\mathbf{x}^\top \mathbf{x}}, \qquad \mathbb{P}\Big(\hat{\beta} = \beta\Big) = \mathbb{P}\Big(\mathbf{x}^\top \boldsymbol{\epsilon} = 0\Big) = 0$$

since X is a continuous random variable. However, for large enough sample size, it holds (with high probability)

$$i \in \arg \min_l |\hat{\epsilon}_i - \epsilon_l| \ \forall i,$$

i.e., the estimated residuals scatter closely around the true value from the discrete set. There are roughly n/k observations per possible value of $\mathcal{E}$, and, due to the linear dependence,

around each value, the ordering of $\hat{\boldsymbol{\epsilon}}$ corresponds to the ordering of $\mathbf{x}$ or is exactly inverted. Therefore,

$$\hat{R}_i \mod \frac{n}{k} \approx \hat{R}_{M(i)} \mod \frac{n}{k} \text{ and } \hat{R}_{M(i)}|\hat{R}_i \dot{\sim} \hat{R}_i \mod \frac{n}{k} + \frac{n}{k}\text{Unif}\{0, \ldots, k-1\}.$$

Since the $\hat{\epsilon}_i$ all have distinct values, it holds

$$\sum_{i=1}^{n} \hat{L}_i = \sum_{i=1}^{n} i = \frac{n^2 + n}{2} \text{ and} \sum_{i=1}^{n} \hat{L}_i^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \text{ such that}$$

$$T_n(\hat{\boldsymbol{\epsilon}}, \mathbf{x}) = \frac{n \sum_{i=1}^{n} \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} - \frac{n(n+1)(2n+1)}{6}}{\frac{n^3 + n^2}{2} - \frac{n(n+1)(2n+1)}{6}}.$$

We consider the only random term

$$\mathbb{E}\left[\sum_{i=1}^{n} \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}\right] = \mathbb{E}\left[\sum_{\hat{R}_i=1}^{n} \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}\right] =$$

$$\mathbb{E}\left[\sum_{\hat{R}_i=1}^{n} \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}|\hat{R}_1, \ldots, \hat{R}_n\right] = \sum_{\hat{R}_i=1}^{n} \mathbb{E}\left[\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}|\hat{R}_1, \ldots, \hat{R}_n\right] =$$

$$\sum_{\hat{R}_i=1}^{n} \mathbb{E}\left[\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}|\hat{R}_i\right].$$

The first equality holds as summing over all i is the same as summing over all ranks. As the problem is permutation invariant, conditioning on all ranks does not change the expectation. Under the conditioning, the ranks are deterministic and linearity of expectation applies. Finally, knowing any rank apart from $\hat{R}_i$ does not influence $\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}$. We analyse the expectation under the approximate conditional distribution as given above. If $\hat{R}_i \leq n/k$, $\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} = \hat{R}_i$. If $n/k < \hat{R}_i \leq 2n/k$ and the uniformly chosen number is 0, $\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} = \hat{R}_i - n/k$. This has probability $1/k$. Otherwise, $\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} = \hat{R}_i$. Analogously, if $ln/k < \hat{R}_i \leq (l+1)n/k$ for some integer $0 \leq l < k$, i.e., $l = \max\left\{r \in \mathbb{N}_0 | r < \hat{R}_i k/n\right\}$, it holds under the approximate distribution

$$\tilde{\mathbb{E}}\left[\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}|\hat{R}_i\right] = \hat{R}_i - \frac{1}{k}\sum_{r=0}^{l} \frac{rn}{k} = \hat{R}_i - \frac{n}{2k^2}\left(l^2 + l\right).$$

For each possible value of l, there are n/k ranks such that $ln/k < \hat{R}_i \leq (l+1)n/k$. Therefore,

$$\sum_{\hat{R}_i=1}^{n} \tilde{\mathbb{E}}\left[\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}|\hat{R}_i\right] = \sum_{\hat{R}_i=1}^{n} \hat{R}_i - \frac{n^2}{2k^3}\sum_{l=0}^{k-1}\left(l^2 + l\right) =$$

$$\frac{n^2+n}{2} - \frac{n^2}{2k^3}\left(\frac{k(k-1)(2k-1)}{6} + \frac{k(k-1)}{2}\right) = \frac{n^2+n}{2} - \frac{n^2}{6k^3}\left(k^3 - k\right).$$

Then,

$$\mathbb{E}[T_n(\hat{\boldsymbol{\epsilon}}, \mathbf{x})] \approx \frac{\frac{n^3+n^2}{2} - \frac{n(n+1)(2n+1)}{6} - \frac{n^3}{6k^3}\left(k^3-k\right)}{\frac{n^3+n^2}{2} - \frac{n(n+1)(2n+1)}{6}} \xrightarrow{n\to\infty} \frac{\frac{1}{6} - \frac{1}{6} + \frac{1}{6k^2}}{\frac{1}{6}} = \frac{1}{k^2}.$$

To make the proof complete the proof, we need to show that

$$\left|\sum_{\hat{R}_i=1}^{n} \tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big]\right| = \mathcal{O}(n^2).$$

We even control

$$\sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big]\right|.$$

For arbitrary conditioning events A, we have

$$\sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big]\right| =$$

$$\sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i, A\Big]\mathbb{P}(A) -$$

$$\mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i, A^c\Big]\mathbb{P}(A^c)\right| \leq$$

$$\sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i, A\Big]\right|\mathbb{P}(A) +$$

$$\sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i, A^c\Big]\right|\mathbb{P}(A^c) \leq$$

$$\sum_{\hat{R}_i=1}^{n} n\mathbb{P}(A) + \sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i, A^c\Big]\right| =$$

$$\sum_{\hat{R}_i=1}^{n} \left|\tilde{\mathbb{E}}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i\Big] - \mathbb{E}\Big[\min\big\{\hat{R}_i, \hat{R}_{M(i)}\big\}|\hat{R}_i, A^c\Big]\right| + n^2\mathbb{P}(A).$$

Hence, we can ignore events with vanishing probability. Let $v_1, \ldots, v_k$ be the attainable values of $\mathcal{E}$ and

$$n_t = \sum_{i=1}^{n} \mathbb{1}_{\{\epsilon_i = v_t\}} \sim \text{Binom}\left(n, \frac{1}{k}\right).$$

Define the event

$$A = \left\{ \max_t |n_t - n/k| > n^{3/4} \cup \exists i : \ i \notin \arg\min_l |\hat{\epsilon}_i - \epsilon_l| \right\}.$$

By the Markov inequality and a union bound, this event has vanishing probability, so we must only control

$$\sum_{\hat{R}_i = 1}^{n} \left| \tilde{\mathbb{E}} \left[\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} | \hat{R}_i \right] - \mathbb{E} \left[\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} | \hat{R}_i, A^c \right] \right|.$$

Under A^c, there are only $\mathcal{O}(n^{3/4})$ ranks $\hat{R}_i$ for which $\epsilon_i \neq v_{l+1}$ with $l = \max\left\{ r \in \mathbb{N}_0 | r < \hat{R}_i k/n \right\}$ is possible. Summing over these leads to another $\mathcal{O}(n^2)$ term and can be ignored. Consider the $\hat{R}_i$ for which $A_i := \{\epsilon_i = v_{l+1}\}$ holds. Assume without loss of generality that $\hat{\beta} < \beta$ such that larger x_i leads to larger $\hat{\epsilon}_i$. Let $F_X(\cdot)$ be the cumulative distribution function of X. For given $n_1, \ldots, n_k$, we have

$$x_i = F_X^{-1}\left(\frac{\hat{R}_i - \sum_{r=1}^{l} n_r}{n_{l+1}} \right) + \mathcal{O}_p\left(n^{-1/2} \right).$$

Thus, one can condition on x_i being in a $n^{-1/4}$ range around the theoretical quantile for any $n_1, \ldots, n_k$ fulfilling A_i. Call this event, whose complementary event has vanishing probability, B_i. It remains to control

$$\mathbb{E}\left[\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} | \hat{R}_i, A^c, A_i, B_i \right] =$$

$$\sum_{r=1}^{k} \mathbb{E}\left[\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} | \hat{R}_i, A^c, A_i, B_i, \epsilon_{M(i)} = v_r \right] \mathbb{P}\left(\epsilon_{M(i)} = v_r | \hat{R}_i, A^c, A_i, B_i \right) =$$

$$\sum_{r=1}^{k} \mathbb{E}\left[\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} | \hat{R}_i, A^c, A_i, B_i, \epsilon_{M(i)} = v_r \right] \left(\mathbb{P}\left(\epsilon_{M(i)} = v_r | \hat{R}_i \right) + o(1) \right) =$$

$$\sum_{r=1}^{k} \mathbb{E}\left[\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} | \hat{R}_i, A^c, A_i, B_i, \epsilon_{M(i)} = v_r \right] \left(\frac{1}{k} + o(1) \right).$$

If $\epsilon_{M(i)} > \epsilon_i$, it holds $\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} = \hat{R}_i$ and we get the right contribution. If $\epsilon_{M(i)} = \epsilon_i$, the conditional expectation is in $\left[\hat{R}_i - 1, \hat{R}_i \right]$, i.e., only a $\mathcal{O}(1)$ deviation. If $v_{m+1} = \epsilon_{M(i)} < \epsilon_i$,

$$\min\left\{ \hat{R}_i, \hat{R}_{M(i)} \right\} = \hat{R}_{M(i)} = \sum_{r=1}^{m} n_r + \sum_{l:\epsilon_l = v_{m+1}} \mathbb{1}_{\left\{ x_l \leq x_{M(i)} \right\}}$$

$$= \sum_{r=1}^{m} n_r + \sum_{l:\epsilon_l = v_{m+1}} \mathbb{1}_{\{x_l \leq x_i\}} + \mathbb{1}_{\left\{ x_{M(i)} > x_i \right\}}.$$

Under the given conditioning, this is

$$\sum_{r=1}^{m} n_r + n_{m+1}\left(\frac{\hat{R}_i - \sum_{r=1}^{l} n_r}{n_{l+1}} + \mathcal{O}(1)\right) + \mathcal{O}(1) =$$

$$\frac{nm}{k} + \mathcal{O}\left(n^{3/4}\right) + \left(\frac{n}{k} + \mathcal{O}\left(n^{3/4}\right)\right)\frac{\hat{R}_i - nl/k + \mathcal{O}\left(n^{3/4}\right)}{n/k + \mathcal{O}\left(n^{3/4}\right)} + \mathcal{O}(n) =$$

$$\frac{nm}{k} + \left(1 + \mathcal{O}\left(n^{-1/4}\right)\right)\left(\hat{R}_i - \frac{nl}{k} + \mathcal{O}\left(n^{3/4}\right)\right) + \mathcal{O}(n) =$$

$$\frac{nm}{k} + \hat{R}_i - \frac{nl}{k} + \mathcal{O}\left(n^{3/4}\right) + \mathcal{O}(n) = \hat{R}_i - \frac{n(l-m)}{k} + \mathcal{O}(n).$$

In summary,

$$\mathbb{E}\left[\min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\}|\hat{R}_i, A^c, A_i, B_i\right] = \hat{R}_i - \frac{1}{k}\sum_{m=0}^{l-1}\frac{n(l-m)}{k} + \mathcal{O}(n) = \hat{R}_i - \frac{1}{k}\sum_{r=0}^{l}\frac{r}{k} + \mathcal{O}(n).$$

Therefore, each term deviates with $\mathcal{O}(n)$ from the approximate expectation. Summing over $\mathcal{O}(n)$ such deviations leads to $\mathcal{O}(n^2)$ as desired. ∎

A.8 Proof of Theorem 5

Due to (A2*), we have

$$\mathbb{E}[Y|\mathbf{X}] = f_{\mathbf{X}_U Y}\left(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right) + g_{\mathbf{X}_U Y}\left(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)\mathbb{E}\left[f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right],$$
$$\mathrm{Var}(Y|\mathbf{X}) = g_{\mathbf{X}_U Y}^2\left(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)$$
$$\left(\mathbb{E}\left[f_{\mathbf{H}Y}^2\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right] - \mathbb{E}\left[f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right]^2\right),$$
$$\mathcal{E} = \frac{f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right) - \mathbb{E}\left[f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right]}{\sqrt{\mathbb{E}\left[f_{\mathbf{H}Y}^2\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right] - \mathbb{E}\left[f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right]^2}}.$$

As in Section A.1,

$$f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus j}\right) \perp \mathbf{X}_U|\mathbf{X}_{-U} \quad \text{and} \quad \mathbb{E}\left[f_{\mathbf{H}Y}\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right] = \perp \mathbf{X}_U|\mathbf{X}_{-U}$$
$$\text{accordingly} \quad \mathbb{E}\left[f_{\mathbf{H}Y}^2\left(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\right)|\mathbf{X}\right] \perp \mathbf{X}_U|\mathbf{X}_{-U} \quad \text{such that} \quad \mathcal{E} \perp \mathbf{X}_U|\mathbf{X}_{-U}.$$

(A1) together with (B1) implies that only terms involving $\mathbf{X}_U$ can be different for the counterfactual; see Section A.1. In particular, $\mathcal{E}$ cannot change. Hence, Y changes from

$$y = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] + \sqrt{\mathrm{Var}(Y|\mathbf{X} = \mathbf{x})}\epsilon \quad \text{to}$$
$$y' = \mathbb{E}\left[Y|\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U}\right] + \sqrt{\mathrm{Var}\left(Y|\mathbf{X}_U = \mathbf{x}'_U, \mathbf{X}_{-U} = \mathbf{x}_{-U}\right)}\epsilon$$

which is as stated in the theorem. ∎

A.9 Proof of Theorem 6

We have the following supporting result.

Lemma 4. Suppose that (B6) - (B8) hold. Then

$$|\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i| = \mathcal{O}_p(1).$$

With Lemma 4 we have replaced Assumption (B3) which is the only missing part to reconstruct the asymptotic results as in Theorems 3 and 4. ■

A.9.1 PROOF OF LEMMA 4

Let

$$V(\mathbf{x}_i) = f_2(\mathbf{x}_i) - f_1^2(\mathbf{x}_i) \quad \text{and} \quad \hat{V}(\mathbf{x}_i) = \hat{f}_2(\mathbf{x}_i) - \hat{f}_1^2(\mathbf{x}_i).$$

Note that

$$\mathbb{P}\left(\frac{1}{V(\mathbf{x}_i)} < \infty\right) = \mathbb{P}(V(\mathbf{x}_i) > 0) = 1 \quad \text{hence} \quad \frac{1}{V(\mathbf{x}_i)} = \mathcal{O}_p(1) \quad \text{likewise}$$

$$\frac{1}{\sqrt{V(\mathbf{x}_i)}} = \mathcal{O}_p(1).$$

Consider the difference

$$\left|V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)\right| = \left|f_2(\mathbf{x}_i) - (f_2(\mathbf{x}_i) + \mathcal{O}_p(1)) - f_1^2(\mathbf{x}_i) + (f_1(\mathbf{x}_i) + \mathcal{O}_p(1))^2\right|$$

$$\leq \mathcal{O}_p(1) + |f_1(\mathbf{x}_i)|\mathcal{O}_p(1) = \mathcal{O}_p(1).$$

In the last equality we use $\mathbb{E}[f_1(\mathbf{x}_i)] = \mathbb{E}[y_i] < \infty$, otherwise regression would not be possible. Hence,

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{V}(\mathbf{x}_i) \leq 0\right) \leq \lim_{n\to\infty} \mathbb{P}\left(\left|V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)\right| \geq V(\mathbf{x}_i)\right) =$$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\left|V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)\right|}{V(\mathbf{x}_i)} \geq 1\right) = 0.$$

Therefore, we will forthcoming condition on $\hat{V}(\mathbf{x}_i)$ being positive which is asymptotically negligible. We now compare the standard deviation and its estimate and consider the event that the difference is either large or not defined. Fix some $\eta > 0$.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0\right)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 | \hat{V}(\mathbf{x}_i) > 0\right) \mathbb{P}\left(\hat{V}(\mathbf{x}_i) > 0\right) +$$

$$\mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 | \hat{V}(\mathbf{x}_i) \leq 0\right) \mathbb{P}\left(\hat{V}(\mathbf{x}_i) \leq 0\right)$$

$$\leq \lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 | \hat{V}(\mathbf{x}_i) > 0\right) + \mathbb{P}\left(\hat{V}(\mathbf{x}_i) \leq 0\right)$$

$$
= \lim_{n \to \infty} \mathbb{P}\left(\left| \sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)} \right| \geq \eta | \hat{V}(\mathbf{x}_i) > 0 \right)
$$

$$
= \lim_{n \to \infty} \mathbb{P}\left(\left| \frac{V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)} + \sqrt{\hat{V}(\mathbf{x}_i)}} \right| \geq \eta | \hat{V}(\mathbf{x}_i) > 0 \right)
$$

$$
\leq \lim_{n \to \infty} \mathbb{P}\left(\left| \frac{V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)}} \right| \geq \eta | \hat{V}(\mathbf{x}_i) > 0 \right)
$$

$$
\leq \lim_{n \to \infty} \mathbb{P}\left(\left| \frac{V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)}} \right| \geq \eta \right) / \mathbb{P}\left(\hat{V}(\mathbf{x}_i) > 0 \right) = 0.
$$

Consider the residuals assuming a positive variance estimate.

$$
|\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i| = \left| \frac{y_i - \hat{f}_i(\mathbf{x}_i)}{\sqrt{\hat{V}(\mathbf{x}_i)}} - \frac{y_i - f_i(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)}} \right| = \left| \frac{\left(y_i - \hat{f}_i(\mathbf{x}_i) \right)\sqrt{V(\mathbf{x}_i)} - (y_i - f_i(\mathbf{x}_i))\sqrt{\hat{V}(\mathbf{x}_i)}}{\sqrt{\hat{V}(\mathbf{x}_i)}\sqrt{V(\mathbf{x}_i)}} \right|
$$

$$
= \left| \frac{\left(f_i(\mathbf{x}_i) - \hat{f}_i(\mathbf{x}_i) \right)\sqrt{V(\mathbf{x}_i)} + (y_i - f_i(\mathbf{x}_i))\left(\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)} \right)}{\sqrt{\hat{V}(\mathbf{x}_i)}\sqrt{V(\mathbf{x}_i)}} \right|
$$

$$
\leq \frac{1}{\sqrt{\hat{V}(\mathbf{x}_i)}} \left(\left| f_i(\mathbf{x}_i) - \hat{f}_i(\mathbf{x}_i) \right| + |\epsilon_i| \left| \sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)} \right| \right).
$$

Arguing similarly as before, we have

$$
\lim_{n \to \infty} \mathbb{P}\left(|\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 \right) = 0.
$$

If we replace $\hat{\epsilon}_i$ by an arbitrary value in case of a nonpositive variance estimate, it holds

$$
\lim_{n \to \infty} \mathbb{P}(|\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i| \geq \eta) = 0. \quad \blacksquare
$$

References

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070–3102, 2021.

Mona Azadkia, Sourav Chatterjee, and Norman Matloff. *FOCI: Feature Ordering by Conditional Independence*, 2021. URL https://CRAN.R-project.org/package=FOCI. R package version 0.1.3.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as Approximations II: A Model-Free Theory of Parametric Regression. *Statistical Science*, 34(4):545 – 565, 2019. doi: 10.1214/18-STS694. URL `https://doi.org/10.1214/18-STS694`.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. *xgboost: Extreme Gradient Boosting*, 2021. URL `https://CRAN.R-project.org/package=xgboost`. R package version 1.4.1.1.

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*, 2023.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21, 2008.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR, 2023.

Matthew H Larson, Luke A Gilbert, Xiaowo Wang, Wendell A Lim, Jonathan S Weissman, and Lei S Qi. Crispr interference (crispri) for sequence-specific control of gene expression. *Nature Protocols*, 8(11):2180–2196, 2013.

Takashi N Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pages 97–106. PMLR, 2021.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188, 1989.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann Publishers, 1988.

Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2009.

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 3–11, 2012.

Jonas Peters. On the intersection property of conditional independence and its application to causal discovery. *Journal of Causal Inference*, 3(1):97–108, 2015.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Niklas Pfister and Jonas Peters. *dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion*, 2019. URL `https://CRAN.R-project.org/package=dHSIC`. R package version 2.1.

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(1):5–31, 2018.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Christoph Schultheiss and Peter Bühlmann. On the pitfalls of gaussian likelihood scoring for causal discovery. *Journal of Causal Inference*, 11(1), 2023.

Christoph Schultheiss, Peter Bühlmann, and Ming Yuan. Higher-order least squares: assessing partial goodness of fit of linear causal models. *Journal of the American Statistical Association*, 2023.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Rajen D Shah and Richard J Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

Peter Spirtes. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, pages 278–285. PMLR, 2001.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2000.

Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72, 2023.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8): 8052–8072, 2023.

Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

Sascha Xu, Osman A Mian, Alexander Marx, and Jilles Vreeken. Inferring cause and effect in the presence of heteroscedastic noise. In *International Conference on Machine Learning*, pages 24615–24630. PMLR, 2022.