Diss. ETH No. 30250

On possibilities and impossibilities for causal inference with observational data.

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

CHRISTOPH OLIVER SCHULTHEISS

MSc ETH in Data Science, ETH Zurich

born on 15.07.1994

accepted on the recommendation of

Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Niklas Pfister, co-examiner

2024

Acknowledgments

First of all, I want to thank my supervisor and primary collaborator Peter Bühlmann for giving me this opportunity and the support over the years. I received a lot of freedom to pursue the projects that interest me but I could always rely on constructive feedback and input when I needed it. The way he meets doctoral students at eyelevel and recognises that our PhD is mainly for us and not to boost his own profile is truly exceptional. This allowed for a very enjoyable and productive collaboration.

I am thankful to all of the Seminar for Statistics for this great work environment. Coming to work always felt a bit like going to see my friends as well. This motivated me on the days when research itself was maybe not so motivating. Among the many great people I was fortunate to meet here over these years, I want to particularly mention Felix Kuchelmeister and Tanja Finger. I am also grateful to the administrative staff within our group and across the department, including ISG, for being a reliable source of support and information. Adapting the words of JK Rowling: You will also find that help will always be given ~~at Hogwarts~~ in HG to those who ask for it.

Last but surely not least, a special thank you goes to my parents, Claudia and Luc, for always having my back - when I deserved it and when I didn't. Being able to focus on my education without any greater worries was a key to my successful studies. Without that, I might never have ended up in this comfortable position, which a PhD in statistics at ETH is.

The projects to be presented received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 786461).

Christoph Schultheiss
Zurich, October 2024

Abstract

Large datasets, which become increasingly available, allow us to infer interactions between different quantities of interest. However, one of the most fundamental truths in statistics is that "correlation does not imply causation". If we observe strong dependence between two variables, we still do not know anything about their causal relation, i.e., does X cause Y, does Y cause X, or is neither true, and the dependence is induced by a common cause? The gold standard to answer such questions is to design and conduct a randomized experiment, where we have control over one of the variables. If the data still shows dependence between the two, this controlled variable must be a cause.

Unfortunately, such experimental data is often unavailable in practice: it could be infeasible or unethical to collect, or data collection simply happened before the research questions of interest were defined. We call such data observational meaning that it does not come from a designed experiment. In this thesis, we discuss conditions under which we can still infer some causal interpretation from purely observational data, and we provide suitable estimation algorithms.

A promising tool to model multivariate systems are structural equation models also referred to as structural causal models. There, one assumes that an effect variable can be modelled as a function of its causes, which are other variables in the system, and some independent unobservable noise. In most generality, these models do not render the causal connections identifiable: even if we collect infinite amounts of data one cannot decide between different candidate models. But, under suitable restrictions on the function class, i.e., the way the causes and noise interact to yield the effect, they become identifiable.

In Chapter 1, we consider the simplest function class: the linear structural equation model, where the effect is a linear combination of its causes and the noise. This is known to be identifiable if the noise is non-Gaussian. We provide a novel technique called "ancestor regression" which infers the causes of a given target variable of interest. It uses a simple trick and is computationally very easy. Unlike existing methods, we provide asymptotic type I error guarantees against false causal claims, and the error control also works well for finite samples. These guarantees even hold in unidentifiable settings. We extend the method to time series data with linear causal relations in Chapter 2. After slightly adapting the algorithm to this scenario, we can provide similar guarantees. Given the amount of time series datasets in practice, this modification can yield a large increase in scope.

These methods, like many others, rely on the correctness of the assumed model class. In Chapter 3, we bring this into question. We focus on a given assumed linear causal model and assess whether this is plausible. Our well-specification test considers dependencies in higher moments which are present in case of misspecification. It is constructed in a way that allows for a per-covariate statement, i.e., even if there is evidence that the overall model is not true, we could retain some causal interpretation. We search for a subset of the predictors for which linear causal effects are compatible with the data.

We extend these ideas to a less restrictive framework in Chapter 4. In the additive noise model, we assume that the unobserved noise additively disturbs the effect of the causes without constraining this effect further. We design a framework to assess the well-specification of this additive noise model, which similarly allows for an interpretation for subsets of the predictors. Due to the unconstrained nature of modeling the causal effects, it requires a nonparametric assessment of (conditional) independence relations.

We consider a specific misspecification of the additive noise model in Chapter 5. Namely, what happens if one wrongly relies on the additive noise being Gaussian? In some settings assuming Gaussianity is known to be conservative, meaning that one may not exploit all available information and in the worst case remains indecisive. We show that this can be a fallacy in slight variations of these settings, i.e., wrongly relying on the noise's distribution can lead to false causal claims.

In both Chapters 4 and 5, we discuss extensions to heteroskedastic models in which the noise intensity can depend on the causes. More flexibility in the function class could render the model unidentifiable. Therefore, we analyze the identifiability after this extension in Chapter 6.

Zusammenfassung

Große Datensätze, die in zunehmendem Maße zur Verfügung stehen, ermöglichen es uns, Wechselwirkungen zwischen verschiedenen interessanten Größen zu erkennen. Eine der grundlegendsten Wahrheiten in der Statistik ist jedoch: "Korrelation ist nicht gleichbedeutend mit Kausalität". Wenn wir eine starke Abhängigkeit zwischen zwei Variablen beobachten, wissen wir immer noch nichts über ihre kausale Beziehung, d.h. verursacht X Y, verursacht Y X, oder trifft beides nicht zu, und die Abhängigkeit wird durch eine gemeinsame Ursache hervorgerufen? Der Goldstandard zur Beantwortung solcher Fragen ist die Planung und Durchführung eines randomisierten Experiments, bei dem wir eine der Variablen kontrollieren können. Wenn die Daten immer noch eine Abhängigkeit zwischen den beiden Variablen zeigen, muss diese kontrollierte Variable eine Ursache sein.

Leider stehen solche experimentellen Daten in der Praxis oft nicht zur Verfügung: Es könnte undurchführbar oder unethisch sein, sie zu erheben, oder die Datenerhebung erfolgte einfach, bevor die Forschungsfragen von Interesse definiert wurden. Wir bezeichnen solche Daten als Beobachtungsdaten, was bedeutet, dass sie nicht aus einem geplanten Experiment stammen. In dieser Arbeit erörtern wir Bedingungen, unter denen wir aus reinen Beobachtungsdaten dennoch eine gewisse kausale Interpretation ableiten können, und wir stellen geeignete Schätzalgorithmen zur Verfügung.

Ein vielversprechendes Mittel zur Modellierung multivariater Systeme sind Strukturgleichungsmodelle. Sie gehen davon aus, dass eine Effektvariable als Funktion ihrer Ursachen, also anderer Variablen im System, und eines unabhängigen und unbeobachtbaren Rauschens modelliert werden kann. Im allgemeinsten Fall lassen sich mit diesen Modellen die kausalen Zusammenhänge nicht erkennen: Selbst wenn wir unendlich viele Daten sammeln, kann man nicht zwischen verschiedenen in Frage kommenden Modellen entscheiden. Unter geeigneten Einschränkungen hinsichtlich der Funktionsklasse, sprich der Art und Weise, wie die Ursachen und das Rauschen zusammenspielen, um die Wirkung zu erzielen, werden sie jedoch identifizierbar.

In Kapitel 1 betrachten wir die einfachste Funktionsklasse: das lineare Strukturgleichungsmodell, bei dem die Wirkung eine lineare Kombination aus ihren Ursachen und dem Rauschen ist. Es ist bekannt, dass dies identifizierbar ist, wenn das Rauschen nicht normalverteilt ist. Wir stellen eine neuartige Technik namens "Ahnen-Regression" vor, die auf die Ursachen einer bestimmten Zielvariable von Interesse schließen lässt. Sie verwendet einen simplen Trick und ist rechnerisch sehr einfach. Im Gegensatz zu existierenden Methoden bieten wir asymptotische Typ-I Fehlergarantien gegen falsche Kausalaussagen, und die Fehlerkontrolle funktioniert auch für endliche Stichproben gut. Diese Garantien gelten sogar in nicht identifizierbaren Situationen. In Kapitel 2 erweitern wir die Methoden auf Zeitreihendaten mit linearen kausalen Beziehungen. Nach einer leichten Anpassung des Algorithmus an dieses Szenario können wir ähnliche Garantien geben. In Anbetracht der Menge an Zeitreihendaten, die man in der Praxis antrifft, kann diese Erweiterung das Anwedungsgebiet stark

vergrössern.

Diese Methoden beruhen, wie viele andere auch, auf der Korrektheit der angenommenen Modellklasse. In Kapitel 3 stellen wir dies in Frage. Wir konzentrieren uns auf ein gegebenes angenommenes lineares kausales Modell und beurteilen, ob dieses plausibel ist. Unser Spezifikationstest berücksichtigt Abhängigkeiten in höheren Momenten, die im Falle einer Fehlspezifikation vorhanden sind. Er ist so konstruiert, dass er eine Aussage pro Kovariate zulässt, d.h., selbst wenn es Hinweise darauf gibt, dass das Gesamtmodell nicht wahr ist, können wir eine gewisse kausale Interpretation beibehalten. Wir suchen nach einer Teilmenge der Prädiktoren, für die lineare kausale Effekte mit den Daten vereinbar sind.

In Kapitel 4 erweitern wir diese Ideen auf ein weniger restriktives Modell. Im Modell des additiven Rauschens gehen wir davon aus, dass das unbeobachtete Rauschen den Effekt der Ursachen additiv perturbiert, ohne die Form dieses Effekts weiter einzuschränken. Wir entwerfen eine Methode zur Bewertung der korrekten Spezifikation dieses Modells mit additivem Rausch, welche ebenfalls eine Interpretation für Teilmengen der Prädikatoren ermöglicht. Da die Modellierung der kausalen Effekte nicht eingeschränkt ist, ist eine nichtparametrische Bewertung der (bedingten) Unabhängigkeitsbeziehungen erforderlich.

In Kapitel 5 betrachten wir eine spezielle Fehlspezifikation des Modells des additiven Rauschens. Nämlich, was passiert, wenn man sich fälschlicherweise darauf verlässt, dass das additive Rauschen normalverteilt ist? In einigen Szenarien ist die Annahme der Gauß'schen Verteilung als konservativ bekannt, was bedeutet, dass man nicht alle verfügbare Information ausnutzt und im schlimmsten Fall unentschlossen bleibt. Wir zeigen, dass dies in leichten Variationen dieser Szenarien ein Trugschluss sein kann. Fälschlicherweise eine Annahme über die Verteilung des Rauschens zu treffen, kann zu falschen Kausalaussagen führen.

In den beiden Kapiteln 4 und 5 diskutieren wir Erweiterungen auf heteroskedastische Modelle, in denen die Rauschintensität von den Ursachen abhängen kann. Mehr Flexibilität in der Funktionsklasse könnte dazu führen, dass das Modell nicht mehr identifizierbar ist. Daher analysieren wir die Identifizierbarkeit nach dieser Erweiterung in Kapitel 6.

Contents

Outline **x**

1 Ancestor regression in linear structural equation models **1**

1.1 Introduction . . . 2

1.2 Ancestor regression . . . 3

1.2.1 Model and method . . . 3

1.2.2 Adversarial setups . . . 4

1.2.3 Simulation example . . . 6

1.3 Ancestor detection in networks: nodewise and recursive . . . 6

1.3.1 Algorithm and goodness of fit test . . . 6

1.3.2 Simulation example . . . 8

1.4 Real data example . . . 9

1.A Proofs . . . 12

1.A.1 Additional notation . . . 12

1.A.2 Previous work . . . 12

1.A.3 Proof of Theorem 1.1 . . . 12

1.A.4 Proof of Theorem 1.2 . . . 13

1.A.5 Proof of Theoren 1.3 . . . 15

1.A.6 Proof of Proposition 1.1 . . . 16

1.A.7 Proof of Proposition 1.2 . . . 16

1.A.8 Proof of Corollary 1.1 . . . 17

1.B Algorithm . . . 19

1.C Details on the simulation setup . . . 20

1.D Additional simulation results . . . 20

2 Ancestor regression in structural vector autoregressive models **22**

2.1 Introduction . . . 23

2.1.1 Our contribution . . . 23

2.1.2 Structural vector autoregressive model . . . 23

2.1.3 Identifying AN^τ via ancestor regression . . . 24

2.2 Estimation from data and asymptotics . . . 27

2.2.1 Simulation example . . . 28

2.3 Inferring effects in networks . . . 30
2.3.1 Instantaneous effects . . . 30
2.3.2 Summary time graph . . . 30
2.3.3 Simulation example . . . 31
2.4 Real data applications . . . 32
2.5 Discussion . . . 34
2.5.1 Outlook: Lessons for independent data with background knowledge . . . 34
2.5.2 Conclusion . . . 35
2.A Proofs . . . 36
2.A.1 Additional notation . . . 36
2.A.2 Previous work . . . 36
2.A.3 Proof of Theorem 2.1 . . . 36
2.A.4 Proof of Theorem 2.3 . . . 37
2.A.5 Near-epoch dependence . . . 42
2.A.6 Proof of Theorem 2.2 . . . 44
2.A.7 Combined p-values . . . 45
2.B Details on the simulation setup . . . 46

3 Higher-order least squares: assessing partial goodness of fit of linear causal models 48
3.1 Introduction . . . 49
3.1.1 Our contribution . . . 50
3.1.2 Outline . . . 50
3.1.3 Notation . . . 50
3.2 Higher-order least squares (HOLS) . . . 51
3.2.1 Univariate regression as a motivating case . . . 51
3.2.2 Multivariate regression . . . 53
3.2.3 High-dimensional data . . . 57
3.3 The confounded case and local null hypotheses . . . 59
3.3.1 Sample estimates . . . 60
3.3.2 Inferring V from U . . . 62
3.4 Specific models . . . 63
3.4.1 Block independence of $\mathcal{E}_{\mathbf{X}}$ . . . 63
3.4.2 Linear structural equation model . . . 64
3.4.3 Beyond linearity . . . 69
3.5 Real data example . . . 69
3.6 Discussion . . . 71

3.A Simulation results . . . 73
3.A.1 Global null . . . 73
3.A.2 Missing variable in a linear SEM . . . 74
3.A.3 High-dimensional data: missing variable in a linear SEM . . . 75
3.A.4 Confounding onto block-independent $\mathcal{E}_{\mathbf{X}}$ . . . 77
3.B Proofs . . . 80
3.B.1 Proof of Theorem 3.1 . . . 80
3.B.2 Proof of Theorem 3.3 . . . 80
3.B.3 Proof of Theorem 3.4 . . . 83
3.B.4 Proof of Theorem 3.5 . . . 85
3.B.5 Proof of Theorem 3.6 . . . 87
3.B.6 Proof of Theorem 3.8 . . . 89
3.B.7 Proof of Theorem 3.9 . . . 90
3.C Theory for the high-dimensional extension . . . 93
3.C.1 Proof of Lemma 3.8 . . . 96
3.C.2 Proof of Lemma 3.9 . . . 99
3.C.3 Proof of Theorem 3.2 . . . 101

4 Assessing the overall and partial causal well-specification of nonlinear additive noise models 102

4.1 Introduction . . . 103
4.2 Causal well-specification in population . . . 104
4.2.1 Structural causal model . . . 104
4.2.2 Roadmap of our methodology . . . 106
4.2.3 Global well-specification . . . 108
4.2.4 Local well-specification . . . 109
4.3 Estimating the set of well-specified predictor variables . . . 113
4.3.1 Making use of FOCI (Feature Ordering by Conditional Independence) . . . 113
4.3.2 Asymptotic results . . . 115
4.3.3 Practical algorithm . . . 117
4.4 Simulation example . . . 119
4.5 Real data analysis . . . 122
4.6 Location-scale noise models . . . 124
4.6.1 Asymptotic results . . . 125
4.6.2 Simulation example . . . 127
4.7 Conclusion . . . 128

4.A Proofs . . . 129
4.A.1 Proof of Theorem 4.1 . . . 129
4.A.2 Proof of Theorem 4.2 . . . 129
4.A.3 Definitions from FOCI . . . 130
4.A.4 Proof of Proposition 4.1 . . . 131
4.A.5 Proof of Theorem 4.3 . . . 131
4.A.6 Proof of Theorem 4.4 . . . 135
4.A.7 Proof of Proposition 4.2 . . . 136
4.A.8 Proof of Theorem 4.5 . . . 139
4.A.9 Proof of Theorem 4.6 . . . 140

5 On the pitfalls of Gaussian likelihood scoring for causal discovery 143
5.1 Introduction . . . 144
5.2 Data-generating linear model . . . 145
5.2.1 Illustrative examples . . . 147
5.3 Beyond a data-generating linear model . . . 148
5.4 Heteroskedastic noise model . . . 149
5.5 Discussion . . . 151
5.5.1 Data applications . . . 151
5.5.2 Conclusion . . . 152
5.A Proofs . . . 154
5.A.1 Proof of Theorem 5.1 . . . 154
5.A.2 Proof of Proposition 5.1 . . . 154
5.B Derivations for the figures . . . 156
5.B.1 Gaussian and uniform . . . 156
5.B.2 Two Gaussian random variables, or Gaussian and χ^2_1 . . . 156
5.B.3 Two uniform random variables with heteroskedastic fitting . . . 156

6 On the identifiability of causal location-scale noise models 157
6.1 Identifiability of LSNMs . . . 158
6.A Proof . . . 160
6.B Gaussian noise . . . 161

Bibliography 162

Outline

This thesis is divided into six chapters. Chapters 1 to 5 consist, up to minor modifications, of accepted or published journal papers, or preprints submitted to a journal. Chapter 6 is based on a conference paper but shortened to the relevant parts, i.e., the contribution of Christoph Schultheiss.

Chapter 1: Christoph Schultheiss and Peter Bühlmann. Ancestor regression in linear structural equation models. *Biometrika 110 (4), 1117-1124.*

Chapter 2: Christoph Schultheiss and Peter Bühlmann. Ancestor regression in structural vector autoregressive models. *To be revised with minor modifications for publication in Journal of Causal Inference.*

Chapter 3: 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 119 (546), 1019-1031.*

Chapter 4: Christoph Schultheiss and Peter Bühlmann. Assessing the overall and partial causal well-specification of nonlinear additive noise models. *Journal of Machine Learning Research 25, (159): 1-41.*

Chapter 5: Christoph Schultheiss and Peter Bühlmann. On the pitfalls of Gaussian likelihood scoring for causal discovery. *Journal of Causal Inference, 11(1):20220068.*

Chapter 6: Extracted from: 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. *International Conference on Machine Learning, PMLR 202, 14316-14332.*

Supplementary material including proofs, further details on algorithms, and simulations are available in the respective appendices at the end of each chapter.

Chapter 1

Ancestor regression in linear structural equation models

Christoph Schultheiss and Peter Bühlmann
Biometrika 110 (4), 1117-1124.

Abstract

We present a new method for causal discovery in linear structural equation models. We propose a simple "trick" based on statistical testing in linear models that can distinguish between ancestors and non-ancestors of any given variable. Naturally, this can then be extended to estimating the causal order among all variables. We provide explicit error control for false causal discovery, at least asymptotically. This holds true even under Gaussianity, where other methods fail due to non-identifiable structures. These type I error guarantees come at the cost of reduced empirical power. Additionally, we provide an asymptotically valid goodness of fit p-value to assess whether multivariate data stems from a linear structural equation model.

1.1 Introduction

We propose a very simple yet effective method to infer the ancestor variables in a linear structural equation model from observational data.

> Consider a response variable of interest Y and covariates X in a linear structural equation model. The procedure is as follows. For a nonlinear function $f(\cdot)$, for example $f(Y) = Y^3$, run a least squares regression of $f(Y)$ versus Y and all covariates X: the p-value corresponding to the k-th covariate X_k is measuring the significance that X_k is an ancestor variable of Y, and it provides type I error control.

We refer to this method as ancestor regression. Its power (i.e., type II error) depends on the nature of the underlying data-generating probability distribution. Obviously, the proposed method is extremely simple and easy to be used; yet, it deals with the difficult problem of finding the causal order among random variables. In particular, the proposed method does not need any new software and it is computationally very efficient.

Structure search methods based on observational data for the graphical structure in linear structural equation models have been developed extensively for various settings: for the Markov equivalence class in linear Gaussian structural equation models (Spirtes et al., 2000, Chapter 5.4.2; Chickering, 2002) or for the single identifiable directed acyclic graph in non-Gaussian linear structural equation models (Shimizu et al., 2006; Gnecco et al., 2021) or for models with equal error variances (Peters and Bühlmann, 2014). None of the methods comes with p-values and type I error control. In addition, for the identifiable cases, the corresponding algorithms require certain assumptions such as non-Gaussian errors. Particularly, the method from Shimizu et al. (2006) and extensions thereof are not consistent when there are at least two normally distributed additive error terms involved such that false causal claims cannot be avoided even in the large sample limit. If the errors are just slightly non-Gaussian, the method requires very many samples to achieve a favorable behavior. In contrast, our procedure does not rely on any condition apart from linearity, but automatically exploits whether the structure is identifiable or not. In the latter case, we miss out on some causal relationships but our type I error control retains the same asymptotic guarantees. The price to pay for these guarantees is a reduced empirical power compared to competing methods, sometimes being substantial.

Regarding notation, we use upper case letters to denote a random variable, e.g., X or Y. We use lower case letters to denote i.i.d. copies of a random variable, e.g., x. If $X \in \mathbb{R}^p$, then $x \in \mathbb{R}^{n \times p}$. With a slight abuse of notation, x can either denote the copies or realizations thereof. We write x_j to denote the j-th column of matrix x and $x_{i,j}$ to denote the element in row i and column j. With $\leftarrow$, we emphasize that an equality between random variables is induced by a causal mechanism. All proofs are given in Section 1.A in the supplementary material.

1.2 Ancestor regression

1.2.1 Model and method

Let $X \in \mathbb{R}^p$ be given by the following linear structural equation model

$$X_j \leftarrow \Psi_j + \sum_{k \in \mathrm{PA}(j)} \theta_{j,k} X_k \quad j = 1, \ldots, p, \tag{1.1}$$

where the $\Psi_1, \ldots, \Psi_p$ are independent and centered random variables. We assume that $0 < \mathrm{var}(\Psi_j) = \sigma_j^2 < \infty \; \forall j$ such that the covariance matrix of X exists and has full rank. We use the notation $\mathrm{PA}(j)$, $\mathrm{CH}(j)$, $\mathrm{AN}(j)$ and $\mathrm{DE}(j)$ for j's parents, children, ancestors, and descendants. Further, assume that there exists a directed acyclic graph (DAG) representing this structure.

Let X_j with $j \in \{1, \ldots, p\}$ be a variable of interest; it has been denoted as response Y in Section 1.1. Consider a nonlinear function $f(\cdot)$. The following result describes the population property of ancestor regression, with general function $f(\cdot)$.

Theorem 1.1. Assume that the data X follows the model (1.1). Consider the ordinary least squares regression $f(X_j)$ versus X, denote the according OLS parameter by $\beta^{f,j} := \mathbb{E}\big(XX^\top\big)^{-1}\mathbb{E}\{Xf(X_j)\}$ and assume that it exists. Then,

$$\beta_k^{f,j} = 0 \; \forall k \notin \{\mathrm{AN}(j) \cup j\}.$$

Importantly, X_j itself must also be included in the set of predictors. The beauty of Theorem 1.1 lies in the fact that no assumptions on the distribution of the Ψ_l or the size of the $\theta_{l,k}$, apart from existence of the moments, must be taken for any l and $k \in \{1, \ldots, p\}$. This allows one to control against false discovery of ancestor variables.

Typically, $\beta_k^{f,j} \neq 0$ holds for ancestors since a nonlinear function of that ancestor cannot be completely regressed out by the other regressors using only linear terms. For ancestors that are much further upstream, this effect might become vanishingly small. However, this is not such an issue since when fitting a linear model using the detected ancestors, those indirect ancestors are assigned a direct causal effect of 0 anyway.

Based on Theorem 1.1, we suggest testing for $\beta_k^{f,j} \neq 0$ in order to detect some or even all ancestors of X_j. Doing so for all k, requires nothing more than fitting a multiple linear model and using its corresponding z-tests for individual covariates.

Let $x \in \mathbb{R}^{n \times p}$ be n i.i.d. copies from the model (1.1). Define the following quantities

$$\hat{\beta}^{f,j} := \Big(x^\top x\Big)^{-1} x^\top f(x_j), \quad \hat{\sigma}^2 := \frac{\Big\|f(x_j) - x\hat{\beta}^{f,j}\Big\|_2^2}{n-p} \quad \text{and} \quad \widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big) = \Big(x^\top x\Big)^{-1}_{k,k} \hat{\sigma}^2, \tag{1.2}$$

where $f(\cdot)$ is meant to be applied elementwise in $f(x_j)$.

Theorem 1.2. Assume that the data X follows the model (1.1), $\mathbb{E}\Big\{f(X_j)^2\Big\} < \infty$, $\mathbb{E}\big(X_k^4\big) < \infty \; \forall k$ and $\beta^{f,j}$ exists. Let x be n i.i.d copies thereof. Using the definitions from (1.2), it then holds

$$\hat{\beta}_k^{f,j} = \beta_k^{f,j} + \mathcal{O}_p(1), \quad \widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big) = \mathcal{O}_p\left(\frac{1}{n}\right) \quad \text{and}$$
$$z_k^j := \frac{\hat{\beta}_k^{f,j}}{\sqrt{\widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big)}} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \; \forall k \notin \{\mathrm{AN}(j) \cup j\}.$$

Due to this limiting distribution, we suggest testing the null hypothesis $H_{0,k \to j} : \; k \notin \mathrm{AN}(j)$ with the p-value

$$p_k^j = 2\Big\{1 - \Phi\Big(|z_k^j|\Big)\Big\}, \tag{1.3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

For ancestors, for which $\beta_k^{f,j} \neq 0$, the absolute z-statistic increases as $\sqrt{n}$. In typical setups, one can thus detect all ancestors. Having found all ancestors, one could infer the parents with an ordinary least squares regression of X_j versus $X_{\mathrm{AN}(j)}$, using the t-test for assigning the significance of being a parental variable. Such a procedure might have poor error control for low sample sizes as it requires full power in the first step to detect all ancestors; we provide error control only for the estimated ancestral set.

The choice of $f(\cdot)$ has an impact on the constant in the growth of z_k^j for ancestors. If the Ψ_l are symmetric, any even function yields $\beta_k^{f,j} = 0 \; \forall k$. Therefore, odd functions should be used. In our simulations and the real data analysis, we use $f(X_j) = X_j^3$ as it is the simplest odd function that only invokes slightly higher moments than linear functions. This choice leads to empirically competitive performance relative to other candidates in our simulations.

1.2.2 Adversarial setups

There are cases where $\beta_k^{f,j} \neq 0$ does not hold true for some ancestors leading to reduced power of the method. We provide necessary and sufficient conditions for this and present examples. Define first the j-restricted Markov boundary of k to be

$$\mathrm{MA}^{\to j}(k) := \left[\mathrm{PA}(k) \cup \mathrm{CH}(k) \cup \bigcup_{l \in \mathrm{CH}(k)} \{\mathrm{PA}(l) \setminus k\}\right] \cap \{\mathrm{AN}(j) \cup j\}.$$

It contains all the variables in the Markov boundary of k which are ancestors of j or j itself. E.g., if $k \in \mathrm{AN}(j)$ all its parents are in the restricted Markov boundary, but not necessarily all its children.

Theorem 1.3. Let $k \in \mathrm{AN}(j)$. Then,

$$\beta_k^{f,j} = 0 \quad \forall f(\cdot) \quad \text{if and only if} \quad \mathbb{E}(X_k \mid X_j) = \mathbb{E}\Big(X_{\mathrm{MA}^{\to j}(k)}^\top \gamma^{j,k} \mid X_j\Big),$$

where $\gamma^{j,k}$ is the least squares parameter for regressing X_k versus $X_{\mathrm{MA}^{\to j}(k)}$. In particular,

$$\beta_k^{f,j} = 0 \quad \forall f(\cdot) \quad \text{if} \quad \mathbb{E}\Big(X_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big) = X_{\mathrm{MA}^{\to j}(k)}^\top \gamma^{j,k}.$$

Intuitively speaking, if the conditional expectation of X_k given the j-restricted Markov boundary is linear, k could also be a child of all these variables. Thus, it is not detectable as ancestor of j. In the following, we present two examples that fulfil the conditions of Theorem 1.3. These are the only examples we know of.

Gaussian Ψ. It is well-known in causal discovery for linear structural equation models that Gaussian error terms lead to non-identifiability.

Define

$$\mathrm{CH}^{\to j}(k) \coloneqq [\mathrm{CH}(k) \cap \{\mathrm{AN}(j) \cup j\}],$$

i.e., the children of k through which a directed path from k to j begins.

Proposition 1.1. Assume that the data X follows the model (1.1). Let $k \in \mathrm{AN}(j)$ with $\Psi_k \sim \mathcal{N}\big(0, \sigma_k^2\big)$. Then, it holds

$$\beta_k^{f,j} = 0 \quad \forall f(\cdot) \quad \text{if} \quad \Psi_l \sim \mathcal{N}\big(0, \sigma_l^2\big) \ \forall l \in \mathrm{CH}^{\to j}(k).$$

Under the additional assumptions of Theorem 1.2,

$$z_k^j \coloneqq \frac{\hat{\beta}_k^{f,j}}{\sqrt{\widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big)}} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1).$$

Thus, if every directed path from k to j starts with an edge for which the nodes on both ends have Gaussian noise terms, we have no power to detect this ancestor relationship. However, we neither detect the opposite direction as guaranteed by Theorem 1.1, and thus, control against false positives is guaranteed.

Special constellation of distributions and coefficients. A pathological case occurs if a child's, say, l, error term has the same distribution as the inherited contribution from the parent's, say, k, error term. Then, k is not detectable as l's ancestor. Likewise, it is not detected as ancestor of any of l's descendants j to which all directed paths from k start with the edge $k \to l$.

Proposition 1.2. Assume that the data X follows the model (1.1). Let $k \in \mathrm{AN}(j)$ and $\mathrm{CH}^{\to j}(k) = \{l\}$. Then, it holds

$$\beta_k^{f,j} = 0 \quad \forall f(\cdot) \quad \text{if} \quad \Psi_l \overset{\mathbb{D}}{=} \theta_{l,k}\Psi_k.$$

For the variables discussed here, the limiting Gaussian distribution as stated in Theorem 1.2 is not guaranteed even though $\beta_k^{f,j} = 0$; see also the proof in the supplemental material.

1.2.3 Simulation example

We study ancestor regression in a small simulation example. We generate data from a linear structural equation model with 6 variables. The causal order is fixed to be X_1 to X_6. Otherwise, the structure is randomized and changes per simulation run: X_k is a parent of X_l for $k < l$ with probability 0.4 such that there is an average of 6 parental relationships. The edge weights are sampled uniformly and the Ψ_k are assigned by permuting a fixed set of 6 error distributions. The full data generating process can be found in Section 1.C of the supplementary material.

We aim to find the ancestors of X_4 which can be any subset of $\{X_1, X_2, X_3\}$. We create 1000 different setups and test each on sample sizes varying from 10^2 to 10^6. As a nonlinear function, we use $f(X_j) = X_j^3$. By z-statistic, we mean z_k^4 as defined in Theorem 1.2. We calculate p-values according to (1.3) and apply a Bonferroni-Holm correction (without cutting off at 1 for the sake of visualization) to them.

In Figure 1.1, we see the desired $\sqrt{n}$-growth of the absolute z-statistic for the ancestors, while for the non-ancestors their sample averages are close to the theoretical mean under the asymptotic null distribution. Indirect ancestors are harder to detect than parents. Although the null distribution is only asymptotically achieved, the type I family-wise error rate is controlled for every sample size, supporting our method's main benefit, i.e., robustness against false causal discovery.

1.3 Ancestor detection in networks: nodewise and recursive

1.3.1 Algorithm and goodness of fit test

In the previous section, we assumed that there is a (response) variable X_j that is of special interest. This is not always the case. Instead, one might be interested in inferring the full set of causal connections between the variables. Naturally, our ancestor detection technique can be extended to that problem by applying it nodewise. We suggest the procedure sketched below. The detailed algorithm can be found in Section 1.B of the supplementary material. Notably, the algorithm is invariant to the ordering of the variables.

First, the set of ancestors is defined based on the significant p-values, after multiplicity correction over all $p(p-1)$ z-tests, of ancestor regression. Any correction controlling the type I family-wise

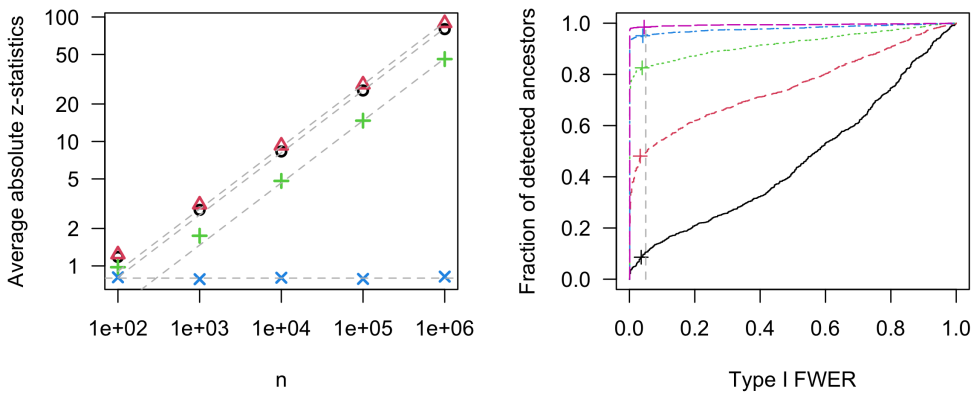


Figure 1.1: Detecting the ancestors of X_4 in a linear structural equation model with 6 variables. The results are based on 1000 simulation runs. On the left: Average absolute z-statistic for all ancestors (circles, black), parents (triangles, red), non-parental ancestors (pluses, green), and non-ancestors (crosses, blue) for different sample sizes. The dashed diagonals correspond to $\sqrt{n}$-growth fitted to perfectly match at $n = 10^5$. The horizontal line corresponds to $(2/\pi)^{1/2}$, i.e., the first absolute moment of the asymptotic null distribution, a standard Gaussian. On the right: fraction of simulation runs with at least one false causal detection versus fraction of detected ancestors for the different sample sizes 10^2 (solid, black), 10^3 (dashed, red), 10^4 (dotted, green), 10^5 (dot-dashed, blue), and 10^6 (long-dashed, pink). The curve uses the level α of the test as implicit curve parameter. The pluses correspond to nominal $\alpha = 5\%$. The vertical line is at actual 5%.

error rate is applicable, and we use here Bonferroni-Holm. Next, further ancestral relationships are constructed recursively by adding the estimated ancestors of every estimated ancestor. This recursive construction facilitates the detection of all ancestors. This procedure cannot increase the type I family-wise error rate compared to just using the significant p-values because a false causal discovery can only be propagated if it existed in the first place.

Since there is no guarantee that the recursive construction does not create directed cycles, i.e., variables are claimed to be their own ancestors, we need to address this. If such cycles are found, the significance level is gradually reduced until no more directed cycles are outputted. This means that the output becomes somewhat independent of the significance level, e.g., in a case with two variables and $p_1^2 = 10^{-6}$ and $p_2^1 = 10^{-3}$ as in (1.3), we would never claim $X_2 \to X_1$ no matter how large α is chosen. We denote the estimated set of ancestors for X_j by $\widehat{\mathrm{AN}}(j)$. Notably, the algorithm determines a causal order between the variables but does not always lead to a unique parental graph. For instance, if $\widehat{\mathrm{AN}}(3) = \{1, 2\}$ and $\widehat{\mathrm{AN}}(2) = \{1\}$, X_1 might be a causal parent of X_3 but its effect could also be fully mediated by X_2.

One can consider the largest significance level such that no loops are created as a p-value for the

null hypothesis that the modeling assumption (1.1) holds true. We denote this level, which is a further output of our algorithm, by $\hat{\alpha}$. Thus, we provide a goodness of fit test for our modelling assumption with an asymptotically valid p-value: a small realized $\hat{\alpha}$ would provide evidence against the linear structural equation model in (1.1). If such evidence exists, it is advisable to take the outcome of ancestor regression or other causal discovery methods relying on linear structural equation models with a grain of salt. We make use of this p-value in the data analysis in Section 1.4. We summarize the properties of our algorithm.

Corollary 1.1. Assume that the conditions of Theorem 1.2 hold $\forall j \in \{1, \dots, p\}$. Let $\widehat{\mathrm{AN}}(j)$ $\forall j \in \{1, \dots, p\}$ and $\hat{\alpha}$ be the output of the nodewise ancestor regression algorithm with significance level α and Bonferroni-Holm correction. Then, it holds

$$\lim_{n\to\infty} \mathbb{P}\Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j) \text{ and } k \in \widehat{\mathrm{AN}}(j)\Big\} \leq \alpha \quad \text{and} \quad \lim_{n\to\infty} \mathbb{P}\big(\hat{\alpha} \leq \alpha'\big) \leq \alpha' \ \forall \alpha' \in (0, \alpha).$$

1.3.2 Simulation example

We extend the simulation from Section 1.2.3 to estimating the ancestors of each variable using the algorithm described in Section 1.3.1. We compare our method to LiNGAM (Shimizu et al., 2006) using the implementation provided in the `R`-package `pcalg` (Kalisch et al., 2012). For every simulation run, we use two slighlty different data generating processes. In the first, only one of the Ψ_k follows a Gaussian distribution, in the second, there are two error terms with normal distribution and an edge between the two respective nodes is always present. As LiNGAM provides an estimated set of parents, we additionally apply our recursive algorithm to the output to get an estimated set of ancestors which enables comparison with our method.

The results are shown in Figure 1.2. For the model with only one Gaussian error variable, we can reliably detect almost all ancestors without any false causal claims for large enough sample sizes. The few exceptions can be explained as some setups can be very close to the non-identifiable case discussed in Proposition 1.2. Not all curves reach a power of 1 even when letting the significance level become arbitrarily large. This can be explained by the possible insensitivity to the significance level, as sketched in Section 1.3.1.

We are able to control the family-wise error rate even for low sample size using a nominal size of $\alpha = 5\%$ supporting our theoretical results. This is not the case for LiNGAM. LiNGAM is designed such that it always must determine a causal order based on the underlying independent component analysis (Hyvarinen, 1999) even when sufficient information is not available. Therefore, no type I error guarantees can be provided. The power of LiNGAM approaches 1 much faster than ancestor regression and if one allows for a bit more liberate type I error, LiNGAM appears preferable in the model with one Gaussian noise term. The picture changes when looking at slight violations of the

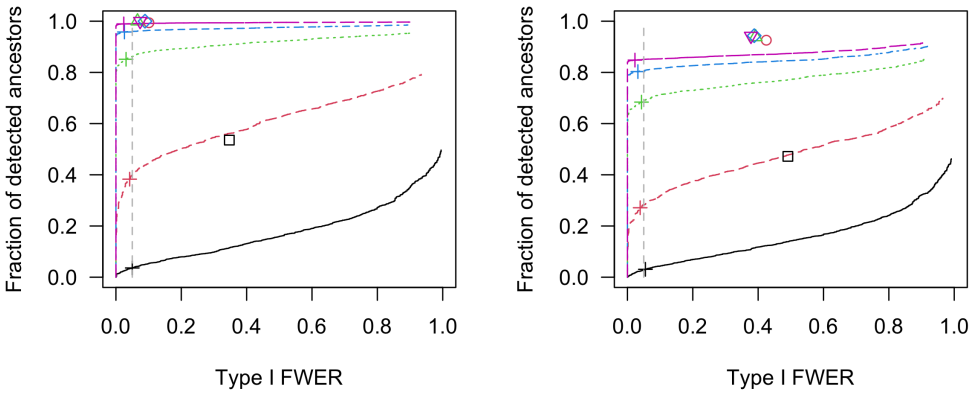


Figure 1.2: Nodewise ancestor detection in a linear structural equation model with 6 variables. The results are based on 1000 simulation runs. Depicted is the family-wise error rate of false causal detection versus the fraction of detected ancestors. The curves use the level of the test α as implicit curve parameter. The pluses correspond to nominal $\alpha = 5\%$. The vertical line is at actual 5%. The other symbols correspond to the performance of the LiNGAM algorithm. We consider the different sample sizes 10^2 (solid / square, black), 10^3 (dashed / circle, red), 10^4 (dotted / triangle pointing upward, green), 10^5 (dot-dashed / diamond, blue), and 10^6 (long-dashed / triangle pointing downward, pink). On the left: exactly 1 error term follows a Gaussian distribution. On the right: exactly 2 error terms follow a Gaussian distribution.

LiNGAM assumption, i.e., another Gaussian error term. LiNGAM is still more powerful but does not control the error at all. No matter the sample size, a wrong causal claim is made in around 40% of the setups. Ancestor regression is more robust to this deviation as the type I error guarantees do not require non-Gaussian error terms. For the unidentifiable edges, it avoids making any decision and can control the error rate at any desired level at the price of some power reduction. In this simulation, Proposition 1.1 applies to around 14% of the ancestral connections.

We provide additional simulation results for settings varying between non-Gaussian and Gaussian scenarios in Section 1.D in the supplementary material. When being close to the fully Gaussian case, despite satisfying the LiNGAM assumption (Shimizu et al., 2006) in population, this clearly worsens the performance of LiNGAM for finite sample size.

1.4 Real data example

We analyze the flow cytometry dataset presented by Sachs et al. (2005). It contains cytometry measurements of 11 phosphorylated proteins and phospholipids. Data is available from various experimental conditions, some of which are interventional environments. The authors provide a "ground truth" on how these quantities affect each other, the so-called consensus network. The dataset has

been further analyzed in various follow-up papers, see, e.g., Mooij and Heskes (2013) and Taeb et al. (2023). Following these works, we consider data from 8 different environments, 7 of which are interventional. The sample size per environment ranges from 707 to 913.

For each environment individually, we estimate the ancestral relationships using our recursive algorithm sketched in Section 1.3.1 with nonlinear function $f(X_j) = X_j^3$ and $\alpha = 0.05$. The goodness of fit p-value $\hat{\alpha}$ per environment, but corrected for the number of environments, ranges from 0.14 to 3×10^{-12}. All but one p-value are lower than 0.04, indicating for these environments that the data does not follow the model (1.1). The deviation can be in terms of hidden variables, nonlinear effects, or noise that is not additive. While the before mentioned and other published findings usually result in one graph harmonized over different environments, our highly varying results across environments suggest to question a standard "autonomy assumption" in causality (Aldrich, 1989) that an intervention does not change the underlying graph (except for edges that point into the intervened node).

Subsequently, we focus on the environment with the highest $\hat{\alpha}$ which seems to be most conformable with a linear structural equation model. The dataset contains 723 observations. For each node, we fit a linear model using the claimed set of ancestors as predictors to see which ancestors might be direct parents. We summarize our findings in Table 1.1. Most ancestors show indication of being direct parents. However, as laid out in Section 1.2.1, we do not have type I error guarantees for finding parents in case some ancestors are missing.

For comparison, we show what conclusion the consensus network as well as Mooij and Heskes (2013) draw for these edges. Our method is in agreement with at least one of these works except for

Causal effect	ancestor regression	linear regression	SC	MH
PIP3 → PIP2	3.3e-39	5.5e-43	$\longrightarrow$	$\longrightarrow$
PIP3 → PLCg	6.7e-39	1.4e-36	$\longrightarrow$	$\dashrightarrow$
PKA → Erk	2.9e-26	7.2e-2	$\longrightarrow$	$\dashrightarrow$
JNK → p38	6.6e-20	2.4e-19	–	–
PKA → Akt	7.2e-20	9.4e-4	$\longrightarrow$	$\longrightarrow$
JNK → PKC	1.2e-16	5.1e-88	$\longleftarrow$	$\longleftarrow$
RAF → MEK	5.4e-15	0	$\longrightarrow$	$\longleftarrow$
PKC → p38	3.1e-13	0	$\longrightarrow$	$\longrightarrow$
Akt → Erk	7.6e-07	0	–	$\longrightarrow$

Table 1.1: Analysis of the dataset by Sachs et al. (2005). The second column reports the raw p-value from ancestor regression, p_k^j, associated with this edge and the third column the raw p-value from the subsequent linear model fit. The rows are ordered by the p-value from ancestor regression from low to high. We present the conclusions of the consensus network in Sachs et al. (2005) (column SC) and the method from Mooij and Heskes (2013) (column MH): the edge is present ($\longrightarrow$), there exists a directed path with the same orientation but no edge ($\dashrightarrow$), the edge is reversed ($\longleftarrow$), there is no directed path (-).

the two edges coming from JNK. One of the indirect paths in Mooij and Heskes (2013) corresponds to the highest p-value in the linear model fit, which is a further agreement. Our outputted ancestral graph, see Figure 1.3, consists of 4 disconnected components. When considering these components individually, we note that the part containing JNK, where we receive somewhat unexpected findings, has the strongest indication of violating the model assumptions in terms of the goodness of fit p-value $\hat{\alpha}$.

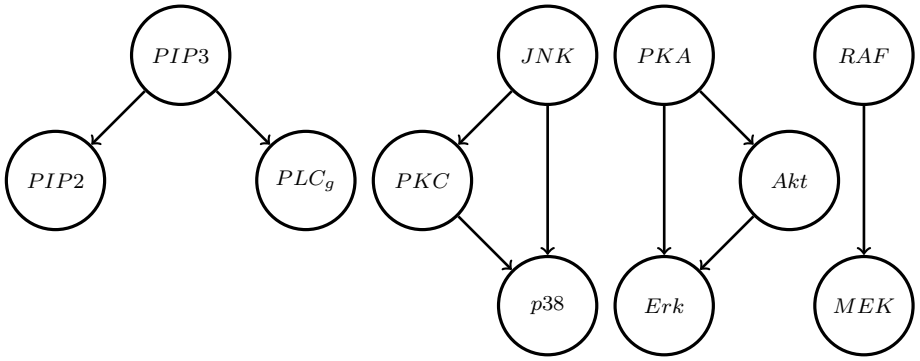


Figure 1.3: Ancestral relations for flow cytometry data obtained with ancestor regression

1.A Proofs

1.A.1 Additional notation

We introduce additional notation that is used for these proofs.

Subindex $-k$, e.g., x_{-k} denotes a matrix with all columns but the k-th. I_n is the n-dimensional identity matrix. P_{-k} denotes the orthogonal projection onto x_{-k} and $P_{-k}^{\perp} = I_n - P_{-k}$ denotes the orthogonal projection onto its complement. P_x is the orthogonal projection onto all x.

For some random vector X, we have the moment matrix $\Sigma^X \coloneqq \mathbb{E}\big(XX^\top\big)$. This equals the covariance matrix for centered X. We assume this matrix to be invertible. Then, the principal submatrix $\Sigma^X_{-j,-j} \coloneqq \mathbb{E}\Big(X_{-j}X_{-j}^\top\Big)$ is also invertible. We denote statistical independence by $\perp$.

1.A.2 Previous work

We adapt some definitions from and results proven in Schultheiss et al. (2024), see also Section 3.2.2.

$$\begin{aligned}
Z_k &\coloneqq X_k - X_{-k}^\top \gamma_k, && \text{where} \\
\gamma_k &\coloneqq \underset{b\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \mathbb{E}\left\{\left(X_k - X_{-k}^\top b\right)^2\right\} = \left(\Sigma^X_{-k,-k}\right)^{-1}\mathbb{E}(X_{-k}X_k), && \\
W_k &\coloneqq f(X_j) - X_{-k}^\top \zeta_k, && \text{where} \\
\zeta_k &\coloneqq \underset{b\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \mathbb{E}\left\{\left(f(X_j) - X_{-k}^\top b\right)^2\right\} = \left(\Sigma^X_{-k,-k}\right)^{-1}\mathbb{E}\{X_{-k}f(X_j)\}. &&
\end{aligned} \tag{1.4}$$

Using these definitions, we have $\beta_k^{f,j} = \mathbb{E}(Z_k W_k)/\mathbb{E}\big(Z_k^2\big) = \mathbb{E}\{Z_k f(X_j)\}/\mathbb{E}\big(Z_k^2\big)$ from partial regression. We cite a Lemma fundamental to our results, see also Lemma 3.5 in Section 3.B.7.

Lemma 1.1. Assume that the data follows the model (1.1) without hidden variables. Then,

$$Z_k = \delta_{k,k}\Psi_k + \sum_{l\in \mathrm{CH}(k)} \delta_{k,l}\Psi_l \quad k = 1,\dots,p$$

for an appropriate set of parameters. Further, the support of γ_j (cf. (1.4)) is restricted to j's Markov boundary.

1.A.3 Proof of Theorem 1.1

Let $\Psi = (\Psi_1,\dots,\Psi_p)^\top$. Then, we can write $X = \omega\Psi$ for a suitable ω with $\omega_{lk} = 0$ if $l \notin \{\mathrm{DE}(k)\cup k\}$. We can now find $\beta^{f,j}$ using this representation.

$$\begin{aligned}
\beta^{f,j} &= \mathbb{E}\Big(XX^\top\Big)^{-1}\mathbb{E}\{Xf(X_j\}) = \left(\omega^{-1}\right)^\top \mathbb{E}\Big(\Psi\Psi^\top\Big)^{-1}\omega^{-1}\omega\mathbb{E}\{\Psi f(X_j)\} \\
&= \left(\omega^{-1}\right)^\top \mathrm{diag}\left\{\frac{1}{\mathbb{E}\big(\Psi_1^2\big)},\dots,\frac{1}{\mathbb{E}\big(\Psi_p^2\big)}\right\}\mathbb{E}\{\Psi f(X_j)\} = \left(\omega^{-1}\right)^\top \left[\frac{\mathbb{E}\{\Psi_1 f(X_j)\}}{\mathbb{E}\big(\Psi_1^2\big)},\dots,\frac{\mathbb{E}\{\Psi_p f(X_j)\}}{\mathbb{E}\big(\Psi_p^2\big)}\right]^\top.
\end{aligned}$$

The third equality follows from the independence of the Ψ_l. Naturally, for all $l \notin \{\text{AN}(j) \cup j\}$ we have $\Psi_l \perp X_j$ such that $\mathbb{E}\{\Psi_l f(X_j)\} = 0$. Further, $\omega_{lk}^{-1} = 0$ if $l \notin \{\text{DE}(k) \cup k\}$. To see this, note that ω would be lower triangular, if $1, \dots, p$ denoted a causal order. Then, its inverse would be lower triangular as well. Naturally, the same principle applies for every other permutation. Thus,

$$\beta_k^{f,j} = \sum_l \omega_{lk}^{-1} \frac{\mathbb{E}\{\Psi_l f(X_j)\}}{\mathbb{E}\big(\Psi_l^2\big)} = \sum_{l \in \{\text{DE}(k) \cup k\}} \omega_{lk}^{-1} \frac{\mathbb{E}\{\Psi_l f(X_j)\}}{\mathbb{E}\big(\Psi_l^2\big)} = \sum_{l \in [\{\text{DE}(k) \cup k\} \cap \{\text{AN}(j) \cup j\}]} \omega_{lk}^{-1} \frac{\mathbb{E}\{\Psi_l f(X_j)\}}{\mathbb{E}\big(\Psi_l^2\big)}$$

such that $\beta_k^{f,j} = 0$ if $\{\text{DE}(k) \cup k\} \cap \{\text{AN}(j) \cup j\} = \emptyset$, i.e., if k is not an ancestor of j.

Alternatively, we could invoke Lemma 1.1 to see that $Z_k \perp X_j$ for a non-ancestor k. Then, $\beta_k^{f,j} = \mathbb{E}\{Z_k f(X_j)\}/\mathbb{E}\big(Z_k^2\big) = 0$

1.A.4 Proof of Theorem 1.2

Define

$$f(X_j) \coloneqq X^\top \beta^{f,j} + \mathcal{E}, \quad \hat{z}_k = P_{-k}^\perp x_k \quad \text{and} \quad \hat{w}_k = P_{-k}^\perp f(x_j) \quad \text{such that} \quad \hat{\beta}_k^{f,j} = \frac{\hat{z}_j^\top \hat{w}_j}{\hat{z}_j^\top \hat{z}_j}.$$

Since we assume the covariance matrix to be bounded, we find

$$\begin{aligned} \frac{1}{n} x_{-k}^\top x_{-k} \xrightarrow{\mathbb{P}} \Sigma_{-k,-k}^X \implies & n\Big(x_{-k}^\top x_{-k}\Big)^{-1} \xrightarrow{\mathbb{P}} \big(\Sigma_{-k,-k}^X\big)^{-1} \\ \implies & \left\| n\Big(x_{-k}^\top x_{-k}\Big)^{-1} \right\| \xrightarrow{\mathbb{P}} \left\| \big(\Sigma_{-k,-k}^X\big)^{-1} \right\| = \mathcal{O}(1), \end{aligned}$$

where we use invertibility and the continuous mapping theorem. This then implies

$$\begin{aligned} |z_k^\top P_{-k} w_k| &= |z_k^\top x_{-k} \Big(x_{-k}^\top x_{-k}\Big)^{-1} x_{-k}^\top w_k| \leq \left\| z_k^\top x_{-k} \right\|_2 \left\| \Big(x_{-k}^\top x_{-k}\Big)^{-1} \right\|_2 \left\| x_{-k}^\top w_k \right\|_2 \\ &\leq \left\| z_k^\top x_{-k} \right\|_1 \left\| \Big(x_{-k}^\top x_{-k}\Big)^{-1} \right\|_2 \left\| x_{-k}^\top w_k \right\|_1 = \sum_{l \neq k} |z_k^\top x_l| \left\| \Big(x_{-k}^\top x_{-k}\Big)^{-1} \right\|_2 \sum_{l \neq k} |x_l^\top w_k| \\ &= \mathcal{O}_p\big(\sqrt{n}\big) \mathcal{O}_p\left(\frac{1}{n}\right) o_p(n) = o_p\big(\sqrt{n}\big) \end{aligned}$$

and analogously

$$|z_k^\top P_{-k} z_k| = \mathcal{O}_p\big(\sqrt{n}\big) \mathcal{O}_p\left(\frac{1}{n}\right) \mathcal{O}_p\big(\sqrt{n}\big) = \mathcal{O}_p(1).$$

We get a better rate for $|z_k^\top x_l|$ than for $|x_l^\top w_k|$ since we assume existence of the fourth moments. Then,

$$\begin{aligned}
\frac{1}{n}\hat{z}_k^\top \hat{w}_k &= \frac{1}{n} z_k^\top P_{-k}^\perp w_k = \frac{1}{n}\Big(z_k^\top w_k - z_k^\top P_{-k} w_k\Big) = \frac{1}{n} z_k^\top w_k + \mathcal{O}_p\Big(\frac{1}{\sqrt{n}}\Big) \\
&= \mathbb{E}(Z_k W_k) + \mathcal{O}_p(1), \\
\frac{1}{\sqrt{n}}\hat{z}_k^\top \hat{w}_k &= \frac{1}{\sqrt{n}} z_k^\top w_k + \mathcal{O}_p(1) \xrightarrow{\mathbb{D}} \mathcal{N}\{\mathbb{E}(Z_k W_k), \mathrm{var}(Z_k W_k)\} \quad \text{and} \\
\frac{1}{n}\hat{z}_k^\top \hat{z}_k &= \frac{1}{n} z_k^\top P_{-k}^\perp z_k = \frac{1}{n}\Big(z_k^\top z_k - z_k^\top P_{-k} z_k\Big) = \frac{1}{n} z_k^\top z_k + \mathcal{O}_p\Big(\frac{1}{n}\Big) \\
&= \mathbb{E}\big(Z_k^2\big) + \mathcal{O}_p(1).
\end{aligned} \tag{1.5}$$

The second line is restricted to covariates for which this variance exists, which includes all non-ancestors as $Z_k \perp W_k$. Using Slutsky's theorem, we have

$$\begin{aligned}
\hat{\beta}_k^{f,j} &= \frac{\mathbb{E}(Z_k W_k)}{\mathbb{E}\big(Z_k^2\big)} + \mathcal{O}_p(1) = \beta_k^{f,j} + \mathcal{O}_p(1)\ \forall k \quad \text{and} \\
\sqrt{n}\hat{\beta}_k^{f,j} &\xrightarrow{\mathbb{D}} \mathcal{N}\left\{0, \frac{\mathbb{E}\big(W_k^2\big)}{\mathbb{E}\big(Z_k^2\big)}\right\} \forall k \notin \{\mathrm{AN}(j) \cup j\},
\end{aligned} \tag{1.6}$$

which proves the first part of the theorem.

It remains to consider the variance estimate. Similar to above, we have

$$n\Big(x^\top x\Big)^{-1}_{kk} \xrightarrow{\mathbb{P}} \big(\Sigma^X\big)^{-1}_{kk} \equiv \frac{1}{\mathbb{E}\big(Z_k^2\big)} = \mathcal{O}(1).$$

Further,

$$\hat{\sigma}^2 = \frac{\Big\|f(x_j) - x\hat{\beta}^{f,j}\Big\|_2^2}{n-p} \coloneqq \frac{1}{n-p}\hat{\epsilon}^\top\hat{\epsilon} = \frac{1}{n-p}\epsilon^\top P_x^\perp \epsilon = \frac{1}{n-p}\Big(\epsilon^\top\epsilon - \epsilon^\top P_x \epsilon\Big).$$

Similar to before

$$\begin{aligned}
\frac{1}{n-p}|\epsilon^\top P_x \epsilon| = \frac{1}{n-p}\mathcal{O}_p(n)\mathcal{O}_p\Big(\frac{1}{n}\Big)\mathcal{O}_p(n) = \mathcal{O}_p(1) \quad \text{such that} \\
\hat{\sigma}^2 = \frac{1}{n-p}\epsilon^\top\epsilon + \mathcal{O}_p(1) = \mathbb{E}\big(\mathcal{E}^2\big) + \mathcal{O}_p(1) = \mathcal{O}_p(1).
\end{aligned}$$

Combined, we find

$$n\widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big) = \mathcal{O}_p(1) \leftrightarrow \widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big) = \mathcal{O}_p\Big(\frac{1}{n}\Big),$$

proving the second part of the theorem.

For non-ancestors, $\beta_k^{f,j} = 0$ such that $W_k = \mathcal{E}$. Then,

$$n\widehat{\mathrm{var}}\Big(\hat{\beta}_k^{f,j}\Big) = \frac{\mathbb{E}\big(\mathcal{E}^2\big)}{\mathbb{E}\big(Z_k^2\big)} + \mathcal{O}_p(1) = \frac{\mathbb{E}\big(W_k^2\big)}{\mathbb{E}\big(Z_k^2\big)} + \mathcal{O}_p(1)$$

such that the last statement of Theorem 1.2 follows again from Slutsky's theorem and (1.6).

1.A.5 Proof of Theoren 1.3

We generally have the following identity

$$\mathbb{E}\{X_l f(X_j)\} = \mathbb{E}\{\mathbb{E}(X_l \mid X_j) f(X_j)\}.$$

Consider first the simpler case where the j-restricted Markov boundary is the full Markov boundary, i.e., all children of k are ancestors of j or j itself. Let $\Omega \coloneqq \mathbb{E}\big(XX^\top\big)^{-1}$ and $\Omega_{kk} \coloneqq d_k$. Then, we have the off-diagonal elements

$$\Omega_{kl} = \begin{cases} -d_k \gamma_l^{j,k} & \text{if } l \in \mathrm{MA}^{\to j}(k) \\ 0 & \text{otherwise} \end{cases},$$

which is a standard fact from least squares regression. Thus,

$$\begin{aligned}
\beta_k^{f,j} &= \sum_{l=1}^{p} \Omega_{kl} \mathbb{E}\{X_l f(X_j)\} = d_k \mathbb{E}\{X_k f(X_j)\} - d_k \sum_{l \in \mathrm{MA}^{\to j}(k)} \gamma_l^{j,k} \mathbb{E}\{X_l f(X_j)\} \\
&= d_k \mathbb{E}\{\mathbb{E}(X_k \mid X_j) f(X_j)\} - d_k \sum_{l \in \mathrm{MA}^{\to j}(k)} \gamma_l^{j,k} \mathbb{E}\{\mathbb{E}(X_l \mid X_j) f(X_j)\} \\
&= d_k \mathbb{E}\left\{ \mathbb{E}\left(X_k - \sum_{l \in \mathrm{MA}^{\to j}(k)} \gamma_l^{j,k} X_l \mid X_j \right) f(X_j) \right\}
\end{aligned}$$

This quantity is 0 for all possible $f(\cdot)$ iff the conditional expectation is the constant 0-function. The if-statement is trivial. For the only if, note that one could choose

$$f(X_j) = \mathbb{E}\left(X_k - \sum_{l \in \mathrm{MA}^{\to j}(k)} \gamma_l^{j,k} X_l \mid X_j \right)$$

leading to a nonzero expectation unless $f(X_j) \equiv 0$. Using

$$\mathbb{E}(X_k \mid X_j) = \mathbb{E}\Big\{ \mathbb{E}\Big(X_k \mid X_j, X_{\mathrm{MA}^{\to j}(k)} \Big) \mid X_j \Big\} = \mathbb{E}\Big\{ \mathbb{E}\Big(X_k \mid X_{\mathrm{MA}^{\to j}(k)} \Big) \mid X_j \Big\},$$

the last part of the theorem follows directly.

For the general case, note that for l in the difference between the Markov boundary and the j-restricted Markov boundary, $\beta_l^{f,j} = 0 \ \forall f(\cdot)$ follows from Theorem 1.1. Thus, the least squares parameter for k and $r \in \mathrm{MA}^{\to j}(k)$ is the same as if these variables did not exist. Therefore, the result for the general case follows directly from the simpler case discussed above.

1.A.6 Proof of Proposition 1.1

Consider the least squares solution when only the variables from the restricted Markov boundary are the predictors. From Lemma 1.1, we get that the residuum, say $\tilde{Z}_k$ is a linear combination of Ψ_k and Ψ_l for $l \in \mathrm{CH}^{\to j}(k)$. For every $r \in \mathrm{MA}^{\to j}(k)$,

$$X_r = \sum_{t \in \{\mathrm{AN}(r) \cup r\}} \omega_{rt} \Psi_t,$$

and, dependence with $\tilde{Z}_k$ could only be induced by

$$\tilde{X}_r = \sum_{t \in \{\mathrm{AN}(r) \cup r\} \cap \left\{\mathrm{CH}^{\to j}(k) \cup k\right\}} \omega_{rt} \Psi_t.$$

By the least squares property, $\tilde{X}_r$ and $\tilde{Z}_k$ are uncorrelated. By the Gaussianity of Ψ_k and Ψ_l, this implies independence. Thus, $\tilde{Z}_k$ is independent from all X_r such that the linear least squares fit is also the conditional expectation. Thus, the sufficient condition from Theorem 1.3 for $\beta_k^{f,j} = 0$ holds.

Due to Gaussianity and Lemma 1.1, Z_k is independent from $\mathrm{CH}^{\to j}(k)$, their descendants, and all its non-descendants. Therefore, it is also independent from

$$\mathcal{E} = W_k = f(X_j) - X_{\{\mathrm{AN}(j) \cup j\} \setminus k} \beta^{f,j}_{\{\mathrm{AN}(j) \cup j\} \setminus k}$$

such that the variance as in (1.5) is consistently estimated.

1.A.7 Proof of Proposition 1.2

Decompose the conditional expectation as

$$\mathbb{E}\Big(X_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big) = \mathbb{E}\left(\Psi_k + \sum_{r \in \mathrm{PA}(k)} \theta_{k,r} X_r \mid X_{\mathrm{MA}^{\to j}(k)}\right) = \mathbb{E}\Big(\Psi_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big) + \sum_{r \in \mathrm{PA}(k)} \theta_{k,r} X_r$$

Recall the definition $\mathrm{CH}^{\to j}(k) = \{l\}$. Then,

$$\begin{aligned} X_l = \mathbb{E}(X_l \mid X_l) = \mathbb{E}\Big(X_l \mid X_{\mathrm{MA}^{\to j}(k)}\Big) &= \mathbb{E}\left(\Psi_l + \theta_{l,k}X_k + \sum_{t\in \mathrm{PA}(l)\setminus k} \theta_{l,t}X_t \mid X_{\mathrm{MA}^{\to j}(k)}\right) \\ &= \mathbb{E}\left(\Psi_l + \theta_{l,k}\Psi_k + \theta_{l,k}\sum_{r\in\mathrm{PA}(k)} \theta_{k,r}X_r + \sum_{t\in \mathrm{PA}(l)\setminus k} \theta_{l,t}X_t \mid X_{\mathrm{MA}^{\to j}(k)}\right) \\ &= \mathbb{E}\Big(\Psi_l + \theta_{l,k}\Psi_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big) + \theta_{l,k}\sum_{r\in\mathrm{PA}(k)} \theta_{k,r}X_r + \sum_{t\in \mathrm{PA}(l)\setminus k} \theta_{l,t}X_t \end{aligned}$$

such that

$$X_l - \theta_{l,k}\sum_{r\in\mathrm{PA}(k)} \theta_{k,r}X_r - \sum_{t\in \mathrm{PA}(l)\setminus k} \theta_{l,t}X_t = \mathbb{E}\Big(\Psi_l + \theta_{l,k}\Psi_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big) = 2\theta_{l,k}\mathbb{E}\Big(\Psi_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big).$$

The last equality follows since $\Psi_l \overset{\mathbb{D}}{=} \theta_{l,k}\Psi_k$ and both random variables depend on the conditioning set on the same way. Therefore, all the terms in $\mathbb{E}\Big(X_k \mid X_{\mathrm{MA}^{\to j}(k)}\Big)$ are linear combination such that the sufficient condition from Theorem 1.3 holds.

However, $Z_k \not\perp X_l$ in general such that $\mathrm{var}(Z_kW_k) = \mathbb{E}\big(Z_k^2\big)\mathbb{E}\big(W_k^2\big)$ is not generally true. Then, the limiting distribution of the estimator is not the same as for non-ancestors; see also (1.5).

1.A.8 Proof of Corollary 1.1

Let j, k be such that $k \in \widehat{\mathrm{AN}}(j)$. This means that there is at least one set $M = \{m_0 = k, m_1, \ldots, m_{t-1}, m_t = j\}$ such that $P^{m_s}_{m_{s-1}} < \alpha \ \forall s \in \{1, \ldots, t\}$, where $t \geq 1$. If $k \notin \mathrm{AN}(j)$ at least one of these must correspond to a false causal discovery, i.e., there is an s such that $P^{m_s}_{m_s-1} < \alpha$ but $m_{s-1} \notin \mathrm{AN}(m_s)$. We conclude

$$\Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j) \text{ and } k \in \widehat{\mathrm{AN}}(j)\Big\} \to \Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j) \text{ and } P^j_k < \alpha\Big\}.$$

Let $r = \sum_{j,k\neq j} 1_{\{H_{0,k\to j} \text{ is true}\}}$ denote the number of true null hypotheses. By the construction of Bonferroni-Holm

$$\Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j) \text{ and } P^j_k < \alpha\Big\} \to \Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j) \text{ and } p^j_k < \alpha/r\Big\}.$$

Let $z^j_k = \hat{\beta}^{f,j}_k / \sqrt{\widehat{\mathrm{var}}\Big(\hat{\beta}^{f,j}_k\Big)}$ as used in Theorem 1.2. We find

$$\lim_{n\to\infty} \mathbb{P}\Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j\} \text{ and } k \in \widehat{\mathrm{AN}}(j)\Big) \leq \lim_{n\to\infty} \mathbb{P}\Big\{\exists j, k \neq j : \ k \notin \mathrm{AN}(j) \text{ and } P^j_k < \alpha\Big\}$$

$$
\begin{aligned}
&\leq \lim_{n\to\infty} \mathbb{P}\Big\{\exists j, k \neq j :\ k \notin \mathrm{AN}(j) \text{ and } p_k^j < \alpha/r\Big\} = \lim_{n\to\infty} \mathbb{P}\Big(\min_{j,k\neq j:\ k\notin \mathrm{AN}(j)} p_k^j < \alpha/r\Big) \\
&\leq \lim_{n\to\infty} \sum_{j,k\neq j:\ k\notin \mathrm{AN}(j)} \mathbb{P}\Big(p_k^j < \alpha/r\Big) = \sum_{j,k\neq j:\ k\notin \mathrm{AN}(j)} \lim_{n\to\infty} \mathbb{P}\Big(p_k^j < \alpha/r\Big) \\
&= \sum_{j,k\neq j:\ k\notin \mathrm{AN}(j)} \lim_{n\to\infty} \mathbb{P}\Big\{\Psi\Big(|z_j^k|\Big) > 1 - \alpha/2r\Big\} = \sum_{j,k\neq j:\ k\notin \mathrm{AN}(j)} \lim_{n\to\infty} \mathbb{P}\Big\{|z_j^k| > \Psi^{-1}(1-\alpha/2r)\Big\} \\
&= \sum_{j,k\neq j:\ k\notin \mathrm{AN}(j)} 1 - \lim_{n\to\infty} \mathbb{P}\Big\{|z_j^k| \leq \Psi^{-1}(1-\alpha/2r)\Big\} = \sum_{j,k\neq j:\ k\notin \mathrm{AN}(j)} \alpha/r = \alpha,
\end{aligned}
$$

which proves the first part of the corollary. The second to last equality uses Theorem 1.2 and the continuous mapping theorem.

As the model (1.1) excludes the possibility of directed cycles, any output of BuildRecursive that contains cycles must include at least one false causal detection. If $\hat{\alpha} < \alpha$, it corresponds to the maximal p-value such that including the corresponding ancestor relationship creates cycles. Therefore, it must hold $\min_{j,k\neq j:\ k\notin \mathrm{AN}(j)} P_k^j \leq \hat{\alpha}$ such that

$$
\lim_{n\to\infty} \mathbb{P}\big(\hat{\alpha} \leq \alpha'\big) \leq \lim_{n\to\infty} \mathbb{P}\Big(\min_{j,k\neq j:\ k\notin \mathrm{AN}(j)} P_k^j \leq \alpha'\Big) \leq \alpha'
$$

using similar arguments as above.

1.B Algorithm

Algorithm 1.1 Nodewise and recursive ancestor detection

Input data $x \in \mathbb{R}^{n \times p}$, significance level $\alpha \in (0, 1)$ and nonlinear function $f(\cdot)$

Output Estimated set of ancestors $\widehat{\mathrm{AN}}(j)\ \forall j \in \{1, \ldots, p\}$, adjusted significance level $\hat{\alpha}$

for $j = 1$ to p **do**
  Calculate $p_k^j\ \forall k \neq j$ using (1.2) and (1.3) *# Calculate the p-values of ancestor regression*
Apply a multiplicity correction to the list of $p_k^j\ \forall j, k \neq j$ denote the corrected p-values by P_k^j
 # Store p-values in a matrix, descendants as rows, ancestors as columns
Define $P \in \mathbb{R}^{p \times p}$ such that $P_{j,k} = P_k^j$ and $P_{j,j} = 1$
$(A,\ \hat{\alpha}) \leftarrow$ FindStructure(P, α)
for $j = 1$ to p **do**
  # Transform binary ancestor matrix to a list of ancestors for each node
  $\widehat{\mathrm{AN}}(j) \leftarrow \{k : A_{j,k} = \texttt{TRUE}\}$
procedure FindStructure($P \in \mathbb{R}^{d \times d}$, α)
  # Define ancestors based on significant p-values
  Define $A \in \mathbb{R}^{d \times d}$ such that $A_{j,k} = \texttt{TRUE}$ if $P_{j,k} < \alpha$ and else $A_{j,k} = \texttt{FALSE}$
  # Recursively complete the ancestral sets such that ancestors' ancestors are ancestors
  $A \leftarrow$BuildRecursive(A)
  $I \leftarrow \{j \in \{1, \ldots, d\} :\ A_{j,j} = \texttt{TRUE}\}$ *# Find nodes that lead to cycles*
  if $I = \emptyset$ **then**
    return $(A,\ \alpha)$ *# If no cycles remain, output the result of the current significance level*
  else
    $\hat{\alpha} \leftarrow \max_{j \in I, k \neq j \in I:\ P_{j,k} < \alpha} P_{j,k}$ *# Otherwise, reduce α to remove at least one edge*
    $(A_{I,I},\ \hat{\alpha}) \leftarrow$ FindStrucure($P_{I,I}$, $\hat{\alpha}$) *# Find structure for variables in cycles*
    $A \leftarrow$BuildRecursive(A) *# Once no more cycles occur, complete the ancestral sets*
    return $(A,\ \hat{\alpha})$
procedure BuildRecursive($A \in \mathbb{R}^{d \times d}$)
  for $j = 1$ to d **do**
    $\widehat{\mathrm{AN}}(j) \leftarrow \{k : A_{j,k} = \texttt{TRUE}\}$ *# Initiate ancestors based on p-values*
  for $j = 1$ to d **do**
    $S \leftarrow \emptyset$ *# Set of ancestors that have been checked, initiated as empty*
    while $\widehat{\mathrm{AN}}(j) \setminus S \neq \emptyset$ **do**
      for $k \in \widehat{\mathrm{AN}}(j) \setminus S$ **do**
        # Add ancestors' ancestors until all are checked
        $\widehat{\mathrm{AN}}(j) \leftarrow \widehat{\mathrm{AN}}(j) \cup \widehat{\mathrm{AN}}(k)$ and $S \leftarrow S \cup K$
    $A_{j,\widehat{\mathrm{AN}}(j)} \leftarrow \texttt{TRUE}$ *# Store to matrix format*
  return A

1.C Details on the simulation setup

We use the following distributions for the Ψ_j: two t_7 distributions, a centered Laplace distribution with scale 1, a centered uniform distribution, and a standard normal distribution. Depending on the scenario, the last error distribution is either uniform or Gaussian. The results in Section 1.2.3 are from the former case. All distributions are normalized to having unit variance. For each simulation run, we randomly permute the distributions to assign them to Ψ_1 to Ψ_6.

We create an edge between the two variables with (potentially) Gaussian error term. The remaining 14 edges $X_k \to X_j$ with $k < j$ are present with probability $5/14$ each such that an average of 6 parental connections exists.

We assign preliminary edge weights uniformly in $[0.5, 1]$. These are further scaled such that for every X_j which is not a source node, the standard deviation of

$$\sum_{k \in \mathrm{PA}(j)} \theta_{j,k} X_k$$

is uniformly chosen from $\left[\sqrt{0}.5, \sqrt{2}\right]$. Thus, the signal-to-noise ratio is between $1/2$ and 2.

To initialize the graph and the weights, we use the function `randomDAG` from the `R`-package `pcalg` (Kalisch et al., 2012) before applying our changes to enforce the constraints.

1.D Additional simulation results

To analyse the effect of close to Gaussian error distributions we consider a further variation of the first the scenario in 1.C. Call the normalized error terms from before Ψ'_j. These are mixed with a standard Gaussian component Ψ''_j such that

$$\Psi_j = \sqrt{1-\gamma}\Psi'_j + \sqrt{\gamma}\Psi''_j \ \forall j.$$

Thus, the Gaussian term causes a fraction γ of the variance. We vary γ from 0, which is the setup from before, to 1 in steps of 0.25. We consider the same performance metrics as in Figure 1.2. For the sake of overview, we restrict ourselves to $n = 10^3$ and $n = 10^4$ in Figure 1.4.

For both LiNGAM and ancestor regression, increasing the amount of Gaussianity leads to a performance drop. Thus, not only fully Gaussian error terms harm these methods. For $\gamma = 0.75$, 10^4 samples are not sufficient to keep the type I error of LiNGAM low. For ancestor regression, nearly Gaussian error distribution leads to a substantial drop in power. However, the type I error remains under control supporting Corollary 1.1. While power considerations are clearly in favor of LiNGAM, especially in easy scenarios, our method leads to fewer but more trustworthy findings in close to unidentifiable scenarios within the class of linear structural equation models.

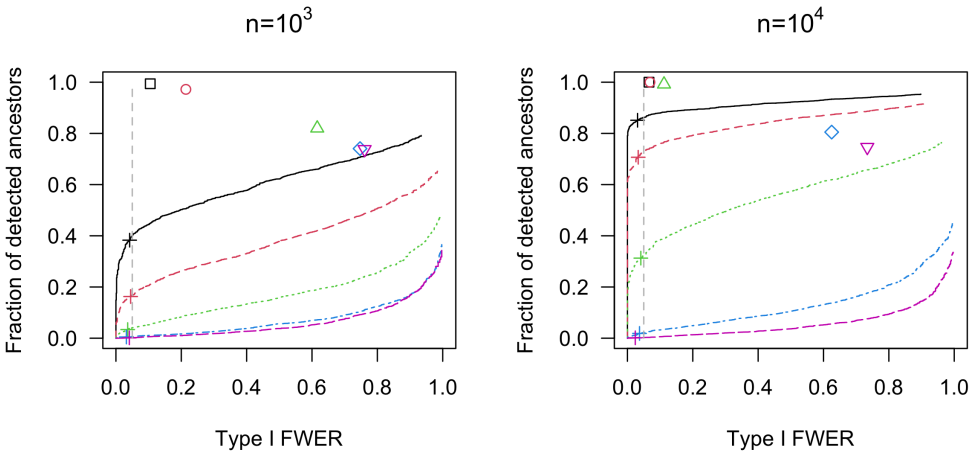


Figure 1.4: Nodewise ancestor detection in a linear structural equation model with 6 variables. The results are based on 1000 simulation runs. Depicted is the family-wise error rate of false causal detection versus the fraction of detected ancestors. The curves use the level of the test α as implicit curve parameter. The pluses correspond to nominal $\alpha = 5\%$. The vertical line is at actual 5%. The other symbols correspond to the performance of the LiNGAM algorithm. We consider the different values of γ: 0 (solid / square, black), 0.25 (dashed / circle, red), 0.5 (dotted / triangle pointing upward, green), 0.75 (dot-dashed / diamond, blue), and 1 (long-dashed / triangle pointing downward, pink). The sample size is 10^3 on the left and 10^4 on the right.

Chapter 2

Ancestor regression in structural vector autoregressive models

Christoph Schultheiss and Peter Bühlmann

To be revised with minor modifications for publication in Journal of Causal Inference.

Abstract

We present a new method for causal discovery in linear structural vector autoregressive models. We adapt an idea designed for independent observations to the case of time series while retaining its favorable properties, i.e., explicit error control for false causal discovery, at least asymptotically. We apply our method to several real-world bivariate time series datasets and discuss its findings which mostly agree with common understanding.

The arrow of time in a model can be interpreted as background knowledge on possible causal mechanisms. Hence, our ideas could be extended to incorporating different background knowledge, even for independent observations.

2.1 Introduction

Real-world datasets often exhibit a time structure violating the i.i.d. assumption widely used in causal discovery and beyond. Such data can be modeled with (structural) vector autoregressive models, i.e., using past and current observations of the time series as predictors. While the time dependence implies certain difficulties in estimation, it offers some advantages as well because a predictor cannot causally affect other variables that represent earlier time points. With independent innovation terms, identifiability guarantees as for fully independent observations can be found under similar structural assumptions, see Peters et al. (2013).

2.1.1 Our contribution

In this work, we extend the recent development on ancestor regression by Schultheiss and Bühlmann (2023a) to the case of multivariate time series with linear causal relations, both instantaneous and lagged. The time dependence between the observations poses technical challenges to ensure the asymptotic guarantees. Further, to obtain error control among the lagged effects, we show how to choose the right adjustment sets for different time lags. Given the amount of time series data encountered in applications, we feel that this extension is of significant practical use; see also the empirical demonstration in Section 2.4.

2.1.2 Structural vector autoregressive model

Let us denote the observed time series by $x_{t,j}$ for $t = 1, \dots, T$ and $j = 1, \dots, d$. At time t the variables are collected to the vector $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,d})^T$. We assume strictly stationary time series, i.e., the probabilistic behavior is the same for every t. We say the time series follows a structural vector autoregressive (SVAR) model of order p if

$$\mathbf{x}_t = \sum_{\tau=0}^{p} \mathbf{B}_\tau \mathbf{x}_{t-\tau} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \dots, \epsilon_{t,d})^T. \tag{2.1}$$

We make the following assumptions:

(A2.1) The $\epsilon_{t,j}$ are centered, independent over both, t and j, and identically distributed over t.

(A2.2) The instantaneous effects in $\mathbf{B}_0$ imply an acyclic structure.

(A2.2) implies that $\mathbf{B}_0$ corresponds to a row- and column-permuted lower triangular matrix. Therefore, the eigenvalues of $I - \mathbf{B}_0$ are all 1 and it is invertible. Hence, we get the equivalent form of our model

$$\mathbf{x}_t = (I - \mathbf{B}_0)^{-1} \left(\sum_{\tau=1}^{p} \mathbf{B}_\tau \mathbf{x}_{t-\tau} + \boldsymbol{\epsilon}_t \right) \coloneqq \sum_{\tau=1}^{p} \tilde{\mathbf{B}}_\tau \mathbf{x}_{t-\tau} + \boldsymbol{\xi}_t, \tag{2.2}$$

where $\boldsymbol{\xi}_t$ has correlated components but is independent over time. Let $\mathbf{x}_{t\_p+1}$ with $p \geq 0$ be the time series in time range t to $t-p$ flattened out to a vector in $\mathbb{R}^{d(p+1)}$. Also $\boldsymbol{\xi}_{t\_p+1}$ are flattened versions of $\boldsymbol{\xi}_t$ patched with zeros, i.e.,

$$\mathbf{x}_{t\_p+1} = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-p} \end{pmatrix} \in \mathbb{R}^{d(p+1)} \quad \text{and} \quad \boldsymbol{\xi}_{t\_p+1} = \begin{pmatrix} \boldsymbol{\xi}_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{d(p+1)}.$$

With these flattened versions, we can rewrite (2.2) as an order 1 model

$$\mathbf{x}_{t\_p} = \begin{pmatrix} \tilde{\mathbf{B}}_1 & \tilde{\mathbf{B}}_2 & \dots & \tilde{\mathbf{B}}_{p-1} & \tilde{\mathbf{B}}_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{pmatrix} \mathbf{x}_{t-1\_p} + \boldsymbol{\xi}_{t\_p} \coloneqq \tilde{\mathbf{B}} \mathbf{x}_{t-1\_p} + \boldsymbol{\xi}_{t\_p}; \tag{2.3}$$

see, e.g., (Lütkepohl, 2005, Chapter 2). We additionally require

(A2.3) The process $\mathbf{x}_t$ is stable, i.e., for $\tilde{\mathbf{B}}$ as in (2.3), $\det\left(I - \tilde{\mathbf{B}}s\right) \neq 0$ if $|s| \leq 1$.

This implies strict stationarity if the process is initialized correctly or has run for an infinite time.

The setting mostly corresponds to the one in Hyvärinen et al. (2010) which extends the LiNGAM method from Shimizu et al. (2006) for linear structural equation models in the i.i.d. setting to the time series case.

Let the τ-lagged causal ancestors of $x_{t,j}$, $\mathrm{AN}^\tau(j)$ be all k for which there exists a directed path from $x_{t-\tau,k}$ to $x_{t,j}$ in the full causal graph. Analogously, we say $k \in \mathrm{PA}^\tau(j) \iff j \in \mathrm{CH}^\tau(k)$ to denote parents and children if there is an edge from $x_{t-\tau,k}$ to $x_{t,j}$. For $\tau = 0$, $\mathrm{AN}^{\tau=0}(j)$ are the instantaneous ancestors of $x_{t,j}$.

2.1.3 Identifying AN^τ via ancestor regression

For the case of linear causal relations in i.i.d. data, the recent development in Schultheiss and Bühlmann (2023a) provides asymptotic type I error guarantees for detecting any covariate's ancestors. The method revolves around the following key observation: Assume that a set of variables x_k, $k \in \{1, \dots, p\}$ is connected by linear causal relations (plus additive noise), and we are interested in the causal ancestors of a given x_j. Then, we can use least squares regression with response variable $f(x_j)$, where $f(\cdot)$ is a nonlinear function, such as $f(x_j) = x_j^3$, and all x_k as predictors, including x_j

itself:

$$f(x_j) \text{ versus } x_j, \{x_k; k \neq j\} \text{ with least squares.}$$

The resulting least squares coefficients are (in population) 0 for all non-ancestors, while they are - up to few counter-examples - non-zero for the ancestors. Hence, one can identify the ancestors using a simple least squares regression. We will show here how this method can be extended to the related SVAR (2.1).

Let

$$\mathbf{x}_t \coloneqq \mathbf{A}_\tau \mathbf{x}_{t-(\tau+1)\_p} + \boldsymbol{\xi}_t^\tau \quad \text{with} \quad \mathbf{x}_{t-(\tau+1)\_p} \perp \boldsymbol{\xi}_t^\tau. \tag{2.4}$$

Here and forthcoming, we use $\perp$ to denote statistical independence. This means that we regress out the contribution of the observations from $\tau + 1$ to $\tau + p$ time steps before. Due to the independence of the innovation terms, such an independent residual can always be found. Hence, $\boldsymbol{\xi}_t^\tau$ are also the corresponding least squares residuals. For $\tau = 0$ and using (2.3), we obtain:

$$\mathbf{A}_0 = \begin{pmatrix} \tilde{\mathbf{B}}_1 & \tilde{\mathbf{B}}_2 & \dots & \tilde{\mathbf{B}}_{p-1} & \tilde{\mathbf{B}}_p \end{pmatrix} \quad \text{and} \quad \boldsymbol{\xi}_t^0 = \boldsymbol{\xi}_t.$$

Define further

$$z_{t,k} \coloneqq \xi_{t,k} - \boldsymbol{\xi}_{t,-k}^\top \boldsymbol{\gamma}_k, \quad \text{where} \quad \boldsymbol{\gamma}_k \coloneqq \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^{d-1}} \mathbb{E}\left[\left(\xi_{t,k} - \boldsymbol{\xi}_{t,-k}^\top \mathbf{b}\right)^2\right] = \mathbb{E}\left[\boldsymbol{\xi}_{t,-k}\boldsymbol{\xi}_{t,-k}^\top\right]^{-1} \mathbb{E}\left[\boldsymbol{\xi}_{t,-k}\xi_{t,k}\right],$$

i.e., the least squares residual of regressing one $\xi_{t,k}$ against all others, or, equivalently, the residual of regressing one $x_{t,k}$ against all others and $\mathbf{x}_{t-1\_p}$.

Theorem 2.1. Assume that $\mathbf{x}_t$ follows the model (2.1) with assumptions (A2.1) to (A2.3). Consider the ordinary least squares regression $f\left(\xi_{t,j}^\tau\right)$ versus $\boldsymbol{\xi}_{t-\tau}$ and denote the according corresponding OLS parameter by

$$\boldsymbol{\beta}^{f,j,\tau} \coloneqq \mathbb{E}\left[\boldsymbol{\xi}_{t-\tau}\boldsymbol{\xi}_{t-\tau}^\top\right]^{-1} \mathbb{E}\left[\boldsymbol{\xi}_{t-\tau} f\left(\xi_{t,j}^\tau\right)\right]$$

and assume that it exists. Then,

$$\beta_k^{f,j,\tau} = \mathbb{E}\left[z_{t-\tau,k} f\left(\xi_{t,j}^\tau\right)\right] / \mathbb{E}\left[z_{t-\tau,k}^2\right] = 0 \quad \forall k \notin \mathrm{AN}^\tau(j).$$

All, $\xi_{t,j}^\tau$ and $\boldsymbol{\xi}_{t-\tau}$, are residuals of a model using $\mathbf{x}_{t-(\tau+1)\_p}$ as predictors and do not depend on time before $t - \tau$.

For $\tau = 0$ this construction corresponds to i.i.d. ancestor regression (Schultheiss and Bühlmann, 2023a) applied to $\boldsymbol{\xi}_t$ which follow an acyclic linear structure equation model as argued in Hyvärinen et al. (2010).

Of particular interest is the reverse statement of Theorem 2.1, namely whether $\beta_k^{f,j,\tau}$ is non-zero

for $k \in \mathrm{AN}^\tau(j)$ for a nonlinear function $f(\cdot)$. While this is typically true, there are some adversarial cases as discussed next.

Adversarial setups

There can be cases where $\beta_k^{f,j,\tau} = 0$ although $k \in \mathrm{AN}^\tau(j)$. For some data generating mechanisms, this happens regardless of the choice of function $f(\cdot)$. We want to characterize these cases. Denote the Markov boundary of $x_{t,k}$ by $\mathrm{MA}(k)$ and get the corresponding vector as

$$\mathbf{x}_{t,\mathrm{MA}(k)} \coloneqq \Big\{ x_{t,l} : l \in \mathrm{CH}^0(k); \quad x_{t-\tau',l} : l \in \mathrm{PA}^{\tau'}(k); \quad x_{t-\tau',l} : l \in \mathrm{PA}^{\tau'}(m) \text{ and } m \in \mathrm{CH}^0(k) \Big\},$$

not including $x_{t,k}$ itself. This matches the classical definition of children, parents, and children's other parents but is restricted to the observed $\mathbf{x}_{t\_p+1}$. Now let

$$\mathrm{CH}^{0,\tau\to j}(k) \coloneqq \mathrm{CH}^0(k) \cap \mathrm{AN}^\tau(j),$$

i.e., the instantaneous children through which a directed path goes to $x_{t+\tau,j}$. Then, we call $\mathrm{MA}^{\tau\to j}(k)$ the restricted Markov boundary and get the corresponding vector as

$$\begin{aligned}\mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)} \coloneqq \Big\{ & x_{t,l} : l \in \mathrm{CH}^{0,\tau\to j}(k); \quad x_{t-\tau',l} : l \in \mathrm{PA}^{\tau'}(k); \\ & x_{t-\tau',l} : l \in \mathrm{PA}^{\tau'}(m) \text{ and } m \in \mathrm{CH}^{0,\tau\to j}(k) \Big\},\end{aligned}$$

again, not including $x_{t,k}$ itself, i.e., only children with a directed path to $x_{t+\tau,j}$ are considered.

Theorem 2.2. Let $k \in \mathrm{AN}^\tau(j)$. Then,

$$\beta_k^{f,j,\tau} = 0 \quad \forall f(\cdot) \quad \text{if and only if} \quad \mathbb{E}\big[x_{t,k} \mid \xi^\tau_{t+\tau,j}\big] = \mathbb{E}\Big[\mathbf{x}^\top_{t,\mathrm{MA}^{\tau\to j}(k)} \boldsymbol{\gamma}^{\tau\to j,k} \mid \xi^\tau_{t+\tau,j}\Big],$$

where $\boldsymbol{\gamma}^{\tau\to j,k}$ is the least squares parameter for regressing $x_{t,k}$ versus $\mathbf{x}_{t,\mathrm{MA}^{\tau\to j}}$. This implies

$$x_{t,k} \perp x_{t+\tau,j} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}.$$

In particular,

$$\beta_k^{f,j,\tau} = 0 \quad \forall f(\cdot) \quad \text{if} \quad \mathbb{E}\Big[x_{t,k} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}\Big] = \mathbf{x}^\top_{t,\mathrm{MA}^{\tau\to j}(k)} \boldsymbol{\gamma}^{\tau\to j,k} \quad \text{and} \quad x_{t,k} \perp x_{t+\tau,j} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}.$$

The implied conditional independence in Theorem 2.2 means that there are no directed paths from $x_{t,k}$ to $x_{t+\tau,j}$ that do not go through some other $x_{t,l}$, or that all these paths cancel each other out. It is trivially always fulfilled for instantaneous effects, i.e., for these, the result is very similar to

Theorem 3 of Schultheiss and Bühlmann (2023a) and the same corresponding examples apply, i.e., Gaussian error terms and identical contribution between predictor and noise, see also Section 1.2.2 for details. Accordingly, if all paths from a lagged ancestor start with an undetectable immediate effect, this lagged effect cannot be detected either.

2.2 Estimation from data and asymptotics

Based on Theorem 2.1, we suggest testing for $\beta_k^{f,j,\tau} \neq 0$ in order to detect some or even all τ-lagged causal ancestors of $x_{t,j}$. Doing so for all k requires nothing more than fitting a multiple linear model and using its corresponding z-tests for individual covariates. Notably, if we are interested in $\mathrm{AN}^\tau(j)$ for several values of τ, we also consider several OLS regressions.

Let

$$\mathbf{x}_{r:s\_p+1} = \begin{pmatrix} \mathbf{x}_{r\_p+1} & \mathbf{x}_{r+1\_p+1} & \dots & \mathbf{x}_{s\_p+1} \end{pmatrix}^\top$$

for some $p+1 \leq r \leq s \leq T$ be a matrix containing predictors at all lags for several time steps. Of course, this matrix has multiple entries corresponding to the same observation. Accordingly,

$$\mathbf{x}_{r:s,j} = \begin{pmatrix} x_{r,j} & x_{r+1,j} & \dots & x_{s,j}. \end{pmatrix}^\top$$

We get the least squares residuals' estimates for the residuals of interest.

$$\begin{aligned}
\hat{\mathbf{z}}_k &= \hat{\mathbf{z}}_{p+1:T,k} && \text{the least squares residual of } \mathbf{x}_{p+1:T,k} \text{ versus } \mathbf{x}_{p+1:T,-k} \text{ and } \mathbf{x}_{p:T-1_p}, \\
\hat{\boldsymbol{\xi}}_k^\tau &= \hat{\boldsymbol{\xi}}_{p+1+\tau:T,k}^\tau && \text{the least squares residual of } \mathbf{x}_{p+1+\tau:T,k} \text{ versus } \mathbf{x}_{p:T-\tau-1_p}, \\
\hat{\boldsymbol{\xi}}^\tau &= \begin{pmatrix} \hat{\boldsymbol{\xi}}_1^\tau & \dots & \hat{\boldsymbol{\xi}}_d^\tau \end{pmatrix}^\top, \\
\hat{\boldsymbol{\xi}} &= \hat{\boldsymbol{\xi}}^0.
\end{aligned}$$

Then, we calculate the following estimates

$$\begin{aligned}
\hat{\beta}_k^{f,j,\tau} &:= \hat{\mathbf{z}}_{p+1:T-\tau,k}^\top f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) / \|\hat{\mathbf{z}}_{p+1:T-\tau,k}\|_2^2 \\
\hat{\sigma}^2 &:= \frac{\left\| f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}_{p+1:T-\tau,k} \hat{\boldsymbol{\beta}}^{f,j,\tau} \right\|_2^2}{T-d-(p+\tau)} \quad \text{and} \\
\widehat{\mathrm{var}}\left(\hat{\beta}_k^{f,j,\tau}\right) &:= \hat{\sigma}^2 / \|\hat{\mathbf{z}}_{p+1:T-\tau,k}\|_2^2,
\end{aligned} \tag{2.5}$$

where $f(\cdot)$ is meant to be applied elementwise in $f\left(\hat{\boldsymbol{\xi}}_j^\tau\right)$. These are the classical least squares estimates for the given predictors and targets. There are d covariates and $T-p-\tau$ observations can be used, hence the given normalization for the variance estimate. We obtain the test statistics

$s_k^{j,\tau} := \dfrac{\hat{\beta}_k^{f,j,\tau}}{\sqrt{\widehat{\text{var}}\Big(\hat{\beta}_k^{f,j,\tau}\Big)}}$ for which we establish asymptotic normality under the null below. Therefore, we suggest testing the null hypothesis

$$H_{0,k^\tau \to j}: \; k \notin \text{AN}^\tau(j)$$

with the p-value

$$p_k^{j,\tau} = 2\Big\{1 - \Phi\Big(|s_k^{j,\tau}|\Big)\Big\}, \tag{2.6}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

To control the asymptotic behavior of these estimates, we make additional assumptions on $f(\cdot)$.

(A2.4) The function $f(\cdot)$ has the following properties

$$\mathbb{E}\Big[f\big(\xi_{t,j}^\tau\big)^2\Big] < \infty \quad \text{and} \quad \exists \delta > 0 \quad \text{such that} \quad \mathbb{E}\Big[\big|\epsilon_{t,k} f\big(\xi_{t+\tau,j}^\tau\big)\big|^{1+\delta}\Big] < \infty \quad \forall k.$$

Also, it is differentiable everywhere, and its derivative $f'(\cdot)$ has the following properties

$$\exists \delta > 0 \quad \text{such that} \quad \mathbb{E}\Big[\big|f'\big(\xi_{t,j}^\tau\big)\big|^{2+\delta}\Big] < \infty \quad \text{and} \quad \mathbb{E}\Big[\big|\epsilon_{t,k} f'\big(\xi_{t+\tau,j}^\tau\big)\big|^{1+\delta}\Big] < \infty \quad \forall k.$$

For monomials of the form

$$f(x) = \text{sign}(x)|x|^\alpha, \quad \alpha > 1,$$

the moment conditions on $f(\cdot)$ imply those on $f'(\cdot)$. We use these functions by default.

Theorem 2.3. Let $\mathbf{x}_t$ follow an SVAR (2.1) for which (A2.1) - (A2.3) hold, and the innovation terms have finite fourth moments. Let $f(\cdot)$ be such that (A2.4) holds and $\boldsymbol{\beta}^{f,j,\tau}$ exists. Using the definitions from (2.5), it then holds for $T \to \infty$

$$\hat{\beta}_k^{f,j,\tau} = \beta_k^{f,j,\tau} + o_p(1), \quad \widehat{\text{var}}\Big(\hat{\beta}_k^{f,j,\tau}\Big) = \mathcal{O}_p\Big(\frac{1}{T}\Big) \quad \text{and}$$
$$s_k^{j,\tau} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \; \forall k \notin \text{AN}^\tau(j).$$

2.2.1 Simulation example

We study ancestor regression in a small simulation example. We generate data from a structural vector autoregressive model with $d = 6$ variables and order $p = 1$. For the instantaneous effects, the causal order is fixed to be $x_{t,1}$ to $x_{t,6}$. Otherwise, the structure is randomized and changes per simulation run: $x_{t,k}$ is an instantaneous parent of $x_{t,l}$ for $k < l$ with probability 0.2 such that there is an average of 3 parental relationships. The edge weights are sampled uniformly and the distributions

of the $\epsilon_{t,k}$ are assigned by permuting a fixed set of 6 error distributions. The entries in $\mathbf{B}_1$ are non-zero with probability 0.1. If so, they are sampled uniformly and assigned a random sign with equal probabilities. If the maximum absolute eigenvalue of $\tilde{\mathbf{B}}$ would be larger than 0.95, $\mathbf{B}_1$ is shrunken such that this absolute eigenvalue is 0.95 to ensure stability.

We aim to find the ancestors of $x_{t,4}$. We create 1000 different setups and test each on sample sizes varying from 10^2 to 10^6. As a nonlinear function, we use $f(x_{t,j}) = x_{t,j}^3$. By z-statistic, we mean $s_k^{4,\tau}$ as in Theorem 2.3. We calculate p-values according to (2.6) and apply a Bonferroni-Holm correction to them.

On the left-hand side of Figure 2.1, we see the average absolute z-statistics for ancestors and non-ancestors for the different sample sizes. We distinguish between three types of ancestors: instantaneous ancestors, lagged ancestors for which $\tilde{\mathbf{B}}_{4,k} \neq 0$, and lagged ancestors from which all causal paths start with an instantaneous effect. The last are the hardest to detect. Otherwise, lagged ancestors have stronger signals than instantaneous ones. This agrees with the intuition that it is easier to find a directed causal path if it is a priori known that only one direction could be possible. For

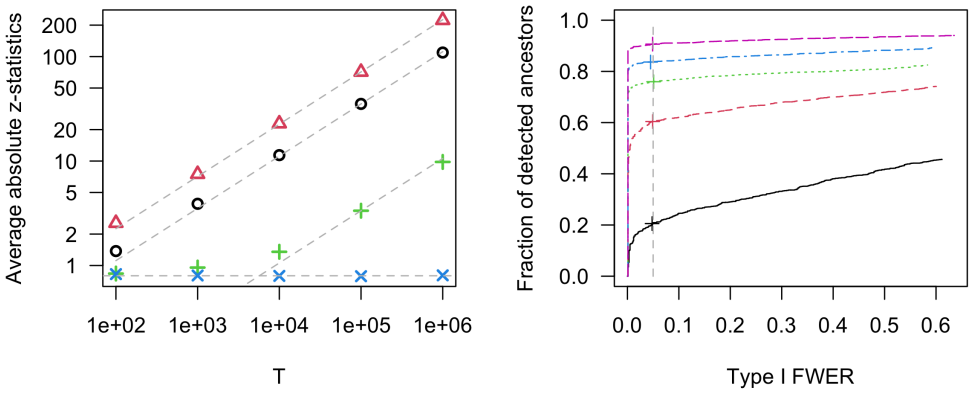


Figure 2.1: Detecting the ancestors of $x_{t,4}$ in a structural vector autoregressive model of order $p = 1$ with 6 variables. The results are based on 1000 simulation runs. On the left: Average absolute z-statistic for instantaneous ancestors (circles, black), lagged ancestors for which $\tilde{\mathbf{B}}_{4,k} \neq 0$ (triangles, red), lagged ancestors from which all causal paths start with an instantaneous effect (pluses, green), and non-ancestors (crosses, blue) for different sample sizes. The dashed diagonals correspond to $\sqrt{T}$-growth fitted to match at $T = 10^5$ perfectly. The horizontal line corresponds to $(2/\pi)^{1/2}$, i.e., the first absolute moment of the asymptotic null distribution, a standard Gaussian. On the right: fraction of simulation runs with at least one false causal detection versus fraction of detected ancestors for the different sample sizes 10^2 (solid, black), 10^3 (dashed, red), 10^4 (dotted, green), 10^5 (dot-dashed, blue), and 10^6 (long-dashed, pink). The curve uses the level α of the test as the implicit curve parameter. The pluses correspond to nominal $\alpha = 5\%$. The vertical line is at actual 5%.

non-ancestors, the observed average of the absolute z-statistics is close to the theoretical mean under the asymptotic null distribution as desired. On the right-hand side, we see that we can control the type I error at the desired level for every sample size. As expected, the power to detect ancestors increases with larger sample sizes. However, driven by this last group of ancestors, there are still some undetected ancestors for $T = 10^6$. For the other groups, we obtain almost perfect power. One could then also infer these missed effects by recursive arguments. We discuss this for the case of networks below.

2.3 Inferring effects in networks

So far, we assumed that there is a time series component $x_{t,j}$ whose causal ancestors are of special interest. This is not always the case. Instead, one might be interested in inferring the full set of causal connections between the variables. Naturally, our ancestor detection technique can be extended to that problem by applying it nodewise. After estimating the effects on every time series, there is a total of $(p+1)d(d-1)$ p-values to consider when ignoring autoregressive effects. We suggest the construction of two types of ancestral graphs that could be of interest.

2.3.1 Instantaneous effects

Focusing on instantaneous effects only, the situation is very similar as in the i.i.d. case discussed in Schultheiss and Bühlmann (2023a). Hence, we apply the same algorithm:

First, we apply a multiplicity correction over the $d(d-1)$ tests to control the type I family-wise error rate. We use the Bonferroni-Holm multiplicity correction. Then, we construct further ancestral relationships recursively: E.g., if $x_{t,1}$ has an instantaneous effect on $x_{t,2}$, and $x_{t,2}$ has an instantaneous effect on $x_{t,3}$, there must be an instantaneous effect from $x_{t,1}$ to $x_{t,3}$. If all detected effects are correct, all such recursively constructed effects must be correct as well. Hence, the type I family-wise error rate remains the same while the power can increase and typically does for larger networks.

If we make type I errors, this could create contradictions leading to cycles. Then, we gradually decrease the significance level for edges within these cycles until no more cycles remain. The largest significance level for which no loops occur is also an asymptotically valid p-value for the null hypothesis that the data come from model (2.2) with assumptions (A2.1) - (A2.4) as we have asymptotic error control under this model class. An example where $\alpha = 0.05$ leads to cycles is demonstrated in Figure 2.2. We keep the edge from $x_{t,2}$ to $x_{t,4}$ using the initial significance level as it is not part of any cycles.

2.3.2 Summary time graph

If not only instantaneous effects but all effects are of interest, one can consider a summary time graph. It includes an edge from k to j if there is a causal path from $x_{t',k}$ to $x_{t,j}$ for any $t' \leq t$. This graph can be cyclic under our modeling assumptions.

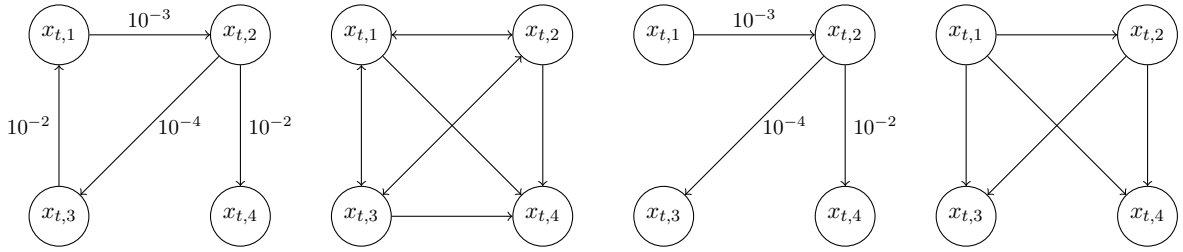


Figure 2.2: From left to right: Detected edges with $\alpha = 0.05$ and the corresponding p-values. Recursive construction leading to cycles. Detected edges with $\alpha < 0.01$ and the corresponding p-values. Recursive construction without cycles.

To obtain it, we first assign a p-value to each potential edge $k \to j$, say, p_k^j. There are $p+1$ p-values corresponding to this edge, i.e., $p_k^{j,0}, \dots p_k^{j,p}$. We combine them using ideas from Meinshausen et al. (2009) designed to combine p-values under arbitrary dependence. For details, see Appendix 2.A.7. Again, we apply the Bonferroni-Holm multiplicity correction to control the family-wise error rate. This allows to add recursively more edges while still controlling the type I error rate. Model (2.1) does not imply that the summary graph must be acyclic. Thus, we output the result after the recursive construction even in the presence of cycles. For example, in Figure 2.2 if the depicted p-values are (multiplicity corrected) summary p-values, the obtained summary graph is the second to the left.

In Figure 2.1, we saw that lagged ancestors from which the directed path begins with an instantaneous edge are the hardest to detect. Here, such a recursive construction can help. Assume $x_{t,1} \to x_{t,2} \to x_{t+1,3}$. Then, there is also a detectable effect from $x_{t,1}$ to $x_{t+1,3}$, but it can be easier to detect the two intermediate edges.

2.3.3 Simulation example

We extend the simulation in 2.2.1 to the network setting. We estimate both, an instantaneous ancestral graph as in Section 2.3.1 and a summary time graph as in Section 2.3.2.

In Figure 2.3, we show the obtained detection rate versus the type I family-wise error rate for varying significance levels. For the summary graph, we obtain slightly better performance. This matches the intuition that this less detailed information is easier to obtain. For either, we achieve essentially perfect separation between ancestors and non-ancestors for large enough sample sizes. Hence, the recursive construction helped to detect even these effects that appeared hard to find based on the individual test as in Figure 2.1. For the instantaneous effects, there is a slight overshoot of the type I error for $T = 100$, i.e., the asymptotic null distribution is not sufficiently attained yet. For all longer time series, it is controlled as desired.

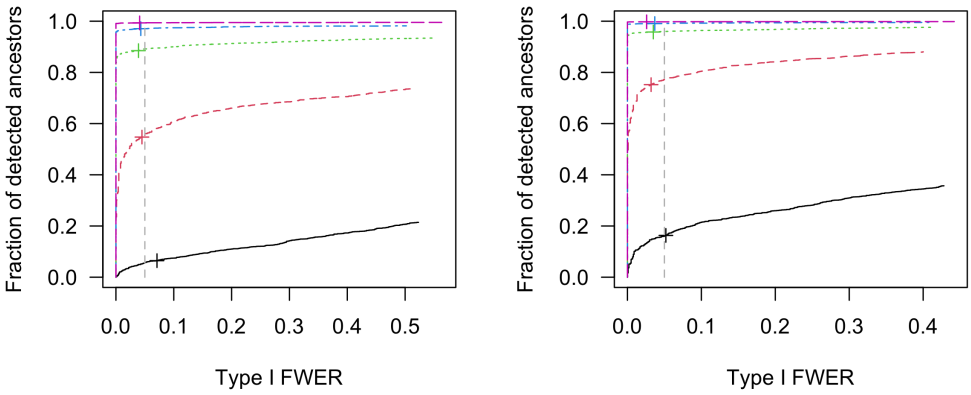


Figure 2.3: Nodewise ancestor detection in a structural vector autoregressive model of order $p = 1$ with 6 variables. The results are based on 1000 simulation runs. Depicted is the family-wise error rate of false causal detection versus the fraction of detected ancestors. The curves use the level of the test α as implicit curve parameter. The pluses correspond to nominal $\alpha = 5\%$. The vertical line is at actual 5%. We consider the different sample sizes 10^2 (solid, black), 10^3 (dashed, red), 10^4 (dotted, green), 10^5 (dot-dashed, blue), and 10^6 (long-dashed, pink). On the left: instantaneous effects. On the right: summary graph.

2.4 Real data applications

Inspired by Peters et al. (2013), we apply our method to several bivariate time series as proof of concept. As they suggest, we fit models of order $p = 6$.

Old Faithful geyser

We analyze data from the Old Faithful geyser (Azzalini and Bowman, 1990) provided in the R-package `MASS` (Venables and Ripley, 2002). It contains information on the waiting time leading to an eruption ($x_{t,1}$) and the duration of an eruption ($x_{t,2}$) for 299 consecutive eruptions. We model these as a bivariate time series although we do not have the classical framework with equidistant measurements in time. The consensus is that the eruption duration affects the subsequent waiting time more than vice-versa.

Our algorithm outputs no significant instantaneous effects on this dataset ($p_2^{1,\tau=0} = 0.78$, $p_1^{2,\tau=0} = 0.73$). This is in line with the consensus which suggests no instantaneous effect from waiting to duration. Here, the duration corresponds to something that happened after the waiting of the corresponding time point such that there should neither be an "instantaneous" effect from duration to waiting. Our summarized p-values suggest an effect from duration to waiting ($p_2^1 = 15 * 10^{-22}$) while the opposite direction is borderline significant ($p_1^2 = 0.094$).

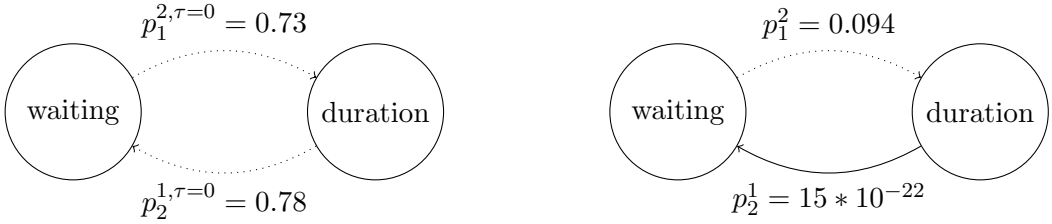


Figure 2.4: Estimated effects for the geyser data. On the left, p-values corresponding to instantaneous effects. On the right, summarized p-values over all considered lags, see Section 2.3.2. Significant edges are drawn as full lines, the others are dotted.

We consider a slightly altered time series shifted such that waiting corresponds to the waiting after the given eruption, i.e., 298 observations remain. Now, our output suggests an instantaneous effect from duration to waiting ($p_2^{1,\tau=0} = 5 * 10^{-4}$, $p_1^{2,\tau=0} = 0.51$) in agreement with the consensus belief. The summarized p-values are $p_2^1 = 9 * 10^{-3}$ and $p_1^2 = 0.18$ respectively.

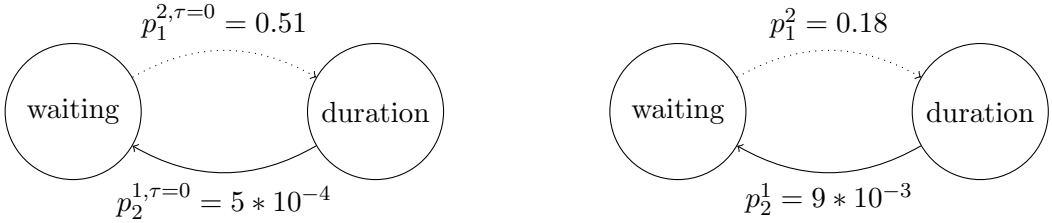


Figure 2.5: Estimated effects for the shifted geyser data. On the left, p-values corresponding to instantaneous effects. On the right, summarized p-values over all considered lags, see Section 2.3.2. Significant edges are drawn as full lines, the others are dotted.

Gas furnace

We look at data from a gas furnace described in Box et al. (2015). It can be downloaded from **https://openmv.net/info/gas-furnace**. The time series are the input gas rate ($x_{t,1}$) and the output CO_2 level observed at 296 equidistant time points. The more plausible causal direction is

Figure 2.6: Estimated effects for the gas furnace data. On the left, p-values corresponding to instantaneous effects. On the right, summarized p-values over all considered lags, see Section 2.3.2. Significant edges are drawn as full lines, the others are dotted.

from input to output.

Our algorithm outputs no instantaneous effects ($p_1^{2,\tau=0} = 0.18$, $p_2^{1,\tau=0} = 0.55$). But, over lags, there appears to be an effect from the input rate to the output concentration as expected. ($p_1^2 = 4 * 10^{-20}$, $p_2^1 = 1$).

Dairy

We use data on ten years of weekly prices for butter ($x_{t,1}$) and cheddar cheese ($x_{t,2}$), i.e., 522 observations in total. Peters et al. (2013) present this as an example where the price of milk could act as a hidden confounder hence violating the model assumptions. Unfortunately, the data source that they mention has disappeared. But, the data was kindly provided by the first author.

We detect no significant instantaneous effects ($p_1^{2,\tau=0} = 0.64$, $p_2^{1,\tau=0} = 0.71$) and hence also no model violations. There is a significant lagged effect from butter to cheddar ($p_1^2 = 5 * 10^{-15}$, $p_2^1 = 1$). In the case of a hidden confounder, both effects should appear in the summary time graph, but we have no evidence for this. However, given the size of the dataset, it can well be that we missed the spurious effect from cheddar to butter.

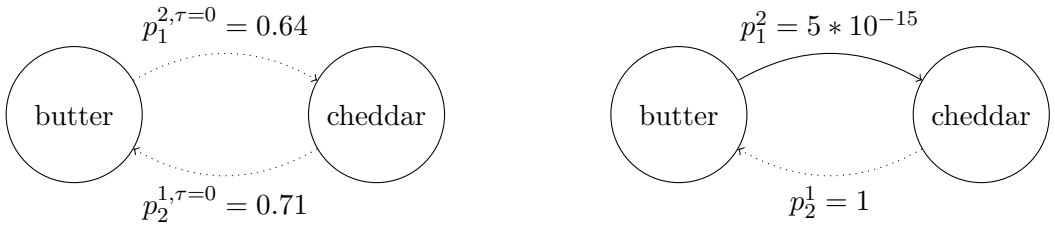


Figure 2.7: Estimated effects for the dairy data. On the left, p-values corresponding to instantaneous effects. On the right, summarized p-values over all considered lags, see Section 2.3.2. Significant edges are drawn as full lines, the others are dotted.

2.5 Discussion

2.5.1 Outlook: Lessons for independent data with background knowledge

Regressing $\xi_{t,j}^{\tau}$ against $\boldsymbol{\xi}_{t-\tau}$ instead of $x_{t,j}$ against $\mathbf{x}_{t-\tau}$ means that we first project out all other covariates that might have a confounding effect but are surely not descendants. Thus, all effects from time points before $t - \tau$ are taken out of the analysis, and the fraction of relevant information in the data increases. If we projected out only after the transformation, i.e., use $f(x_{t,j})$ those effects could not be fully taken out, and the noise level increases.

Similarly, it can happen that for data with no time structure a certain variable is known to be a (potential) confounder between others, but surely not a descendant of any of these. For example, if we measure the weight and height of children, our common sense says that age is a confounder of the

two but surely not causally affected by either. In our assumed framework, i.e., linear causal relations, one can then first regress out this confounding effect which can be done perfectly (in population) before applying the transformation. If it is done after the transformation, there remains some noise terms stemming from the confounder which decreases the signal-to-noise ratio for the effects we are truly interested in.

Importantly, not every variable that is of lesser interest can be regressed out a priori. If its relative place in the causal order is not known, it has to be included in the usual way to retain type I error guarantees. Of course, one can always omit the according tests corresponding to uninteresting effects such that less multiplicity correction must be applied.

2.5.2 Conclusion

We introduce a new method for causal discovery in structural vector autoregressive models. We assess whether there is a causal effect from one time series component to another for any given time lag. The method is computationally very efficient and has asymptotic type I error control against false causal discoveries. Our simulations show that this error control works well for finite time series as well.

We also obtain asymptotic power up to few pathological cases. In networks, additional effects can be inferred by the logic that an ancestor of an ancestor must be an ancestor. In our simulation, we see that this can help to find almost all ancestors without errors even when some connections are individually hard to find at a given sample size.

We apply our method to three real-world bivariate time series and obtain results that mostly agree with the common understanding of the underlying process. Hence, we demonstrate that ancestor regression can be of use even when the modeling assumptions are not fulfilled as in the ideal simulated cases, and the data is only of medium size.

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

2.A Proofs

2.A.1 Additional notation

We introduce additional notation that is used for the proofs. We do not explicitly mention the time steps considered for a regression estimate. It is always meant to use as many observations as available. The number of observations used for an estimate we call simply T as $T \to \infty$ is equivalent to $T - p - \tau \to \infty$.

Subindexing a matrix or vector containing several time lags e.g., $\mathbf{x}_{\_p+1,k}$ means only selecting the column or entry corresponding to time series k with no time lag unless stated otherwise. Subindex $-k$ means all but this column or entry. I_T is the T-dimensional identity matrix. P_{-k} denotes the orthogonal projection onto $\mathbf{x}_{\_p+1,-k}$ and $P_{-k}^{\perp} = I_T - P_{-k}$ denotes the orthogonal projection onto its complement. $P_{\mathbf{x}_{-\tau}}$ is the orthogonal projection onto all $\mathbf{x}_{-\tau-1\_p}$.

For some random vector $\mathbf{x}_{t\_p+1}$, we have the moment matrix $\mathbf{\Sigma}^{\mathbf{x}} \coloneqq \mathbb{E}\big[\mathbf{x}_{t\_p+1}\mathbf{x}_{t\_p+1}^{\top}\big]$. This equals the covariance matrix for centered $\mathbf{x}_{t\_p+1}$. We assume this matrix to be invertible. Then, the principal submatrix $\mathbf{\Sigma}^{\mathbf{x}}_{-j,-j} \coloneqq \mathbb{E}\Big[\mathbf{x}_{t\_p+1,-j}\mathbf{x}_{t\_p+1,-j}^{\top}\Big]$ is also invertible. Again, the negative subindex means the realization without time lag is omitted. We make the analogous assumption for $\mathbf{\Sigma}^{\boldsymbol{\xi}} \coloneqq \mathbb{E}\big[\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{\top}\big]$.

2.A.2 Previous work

We adapt some definitions from and results proved in Schultheiss et al. (2024), see also Section 3.2.2.

$$
\begin{aligned}
z_{t,k} &\coloneqq x_{t,k} - \mathbf{x}_{t\_p+1,-k}^{\top}\boldsymbol{\gamma}_k, && \text{where}\\
\boldsymbol{\gamma}_k &\coloneqq \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^{d(p+1)-1}} \mathbb{E}\Big[\Big(x_{t,k} - \mathbf{x}_{t\_p+1,-k}^{\top}\mathbf{b}\Big)^2\Big] = \big(\Sigma^{\mathbf{x}}_{-k,-k}\big)^{-1}\mathbb{E}[\mathbf{x}_{t\_p+1,-k}x_{t,k}],\\
w_{t,k} &\coloneqq f\big(\xi^{\tau}_{t+\tau,j}\big) - \boldsymbol{\xi}_{t,-k}^{\top}\boldsymbol{\zeta}_k, && \text{where}\\
\boldsymbol{\zeta}_k &\coloneqq \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^{d-1}}\mathbb{E}\Big[\Big(f\big(\xi^{\tau}_{t+\tau,j}\big) - \boldsymbol{\xi}_{t,-k}^{\top}\mathbf{b}\Big)^2\Big] = \Big(\Sigma^{\boldsymbol{\xi}}_{-k,-k}\Big)^{-1}\mathbb{E}\big[\boldsymbol{\xi}_{t,-k}f\big(\xi^{\tau}_{t+\tau,j}\big)\big].
\end{aligned}
\tag{2.7}
$$

We denote by $\mathbf{z}_k \coloneqq \mathbf{x}_{\_p+1,k} - \mathbf{x}_{\_p+1,-k}\boldsymbol{\gamma}_k$ and $\mathbf{w}_k$ analogously these true regression residuals at the relevant time points. For notational simplicity, we do not index $\mathbf{w}_k$ and $\boldsymbol{\zeta}_k$ with the lag τ as the arguments remain the same for every fixed lag. Using these definitions, we have

$$
\beta_k^{f,j,\tau} = \mathbb{E}[z_{t,k}w_{t,k}]/\mathbb{E}\big[z_{t,k}^2\big] = \mathbb{E}\big[z_{t,k}f\big(\xi^{\tau}_{t,j}\big)\big]/\mathbb{E}\big[z_{t,k}^2\big].
$$

from partial regression.

2.A.3 Proof of Theorem 2.1

Under (2.1), $\mathbf{x}_{t\_p+1}$ includes all causal parents of $x_{t,k}$. Now an argument analogous to Lemma 3.5 in Section 3.B.7 shows that $z_{t,k}$ must be a linear combination of $\epsilon_{t,k}$ and possibly some $\epsilon_{t,l}$ where

$k \in \mathrm{AN}^0(l)$. If $k \notin \mathrm{AN}^\tau(j)$, $x_{t+\tau,j}$ must be independent of these innovation terms. Furthermore, $z_{t,k} \perp \mathbf{x}_{t'}\ \forall t' < t$. Hence,

$$z_{t,k} \perp \xi^\tau_{t+\tau,j} = x_{t+\tau,j} - (\mathbf{A}_\tau)_j \mathbf{x}_{t-1\_p},$$

using (2.4). Then,

$$\mathbb{E}\big[z_{t,k} f\big(\xi^\tau_{t+\tau,j}\big)\big] = \mathbb{E}[z_{t,k}]\mathbb{E}\big[f\big(\xi^\tau_{t+\tau,j}\big)\big] = 0$$

Note that as $\beta^{f,j,\tau}_k = 0$, $\boldsymbol{\zeta}_k = \boldsymbol{\beta}^{f,j,\tau}_{-k}$. As k is not a τ-lagged ancestor, its 0-lagged children and descendants cannot be either. Hence, their innovation terms cannot contribute to $\boldsymbol{\xi}^\top_{t,-k}\boldsymbol{\zeta}_k$ such that $z_{t,k} \perp \boldsymbol{\xi}^\top_{t,-k}\boldsymbol{\zeta}_k$ and $z_{t,k} \perp w_{t,k}$.

2.A.4 Proof of Theorem 2.3

Throughout this proof, we apply the law of large numbers in various places. The justification is presented in Section 2.A.5.

With the law of large numbers and the continuous mapping theorem, we get

$$\begin{aligned}
\frac{1}{T}\mathbf{x}^\top_{-p+1}\mathbf{x}_{-p+1} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}^{\mathbf{x}} \implies & \frac{1}{T}\mathbf{x}^\top_{-p+1,-k}\mathbf{x}_{-p+1,-k} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}^{\mathbf{x}}_{-\boldsymbol{k},-\boldsymbol{k}} \\
\implies & T\Big(\mathbf{x}^\top_{-p+1,-k}\mathbf{x}_{-p+1,-k}\Big)^{-1} \xrightarrow{\mathbb{P}} \Big(\boldsymbol{\Sigma}^{\mathbf{x}}_{-\boldsymbol{k},-\boldsymbol{k}}\Big)^{-1} \\
\implies & T\left\|\Big(\mathbf{x}^\top_{-p+1,-k}\mathbf{x}_{-p+1,-k}\Big)^{-1}\right\| \xrightarrow{\mathbb{P}} \left\|\Big(\boldsymbol{\Sigma}^{\mathbf{x}}_{-\boldsymbol{k},-\boldsymbol{k}}\Big)^{-1}\right\| = \mathcal{O}(1)
\end{aligned}$$

For $\mathbf{x}^\top_{-p+1,-k}\mathbf{z}_k/T$ we get a stronger result. Consider any entry

$$\frac{1}{T}\sum_{t=1}^{T} x_{t\_p+1,l} z_{t,k},$$

where l could also represent a time-lagged entry. Now for $t' > t$ consider the autocovariance $\mathbb{E}\big[x_{t\_p+1,l} z_{t,k} x_{t'\_p+1,l} z_{t',k}\big]$. As argued in 2.A.3, $z_{t',k}$ is a combination of innovations from time t' independent of all previous times. Hence, a contribution to the autocovariance could only come from $\mathbb{E}[x_{t\_p+1,l} z_{t,k}]\mathbb{E}\big[x_{t'\_p+1,l} z_{t',k}\big]$. But this is 0 by definition such that there is no time correlation. Combining this with the fourth moment assumption, we can apply the stronger result in Theorem 7.1.1 of Brockwell and Davis (2009) leading to $\left\|\mathbf{x}^\top_{-p+1,-k}\mathbf{z}_k\right\| = \mathcal{O}_p\Big(\sqrt{T}\Big)$. Analogously, as $\boldsymbol{\xi}^\tau_j$ has bounded time dependence $\left\|\mathbf{x}^\top_{-\tau-1\_p}\boldsymbol{\xi}^\tau_j\right\| = \mathcal{O}_p\Big(\sqrt{T}\Big)$. Let $\mathbf{a}_j = (\mathbf{A}_\tau)_j$ be the effect from $\mathbf{x}_{t-1\_p}$ on $x_{t+\tau,j}$ and $\hat{\mathbf{a}}_j$ its least squares estimate.

$$\|\mathbf{z}_k - \hat{\mathbf{z}}_k\|_2^2 = \|P_{-k}\mathbf{z}_k\|_2^2 \leq \left\|\mathbf{z}_k^\top \mathbf{x}_{-p+1,-k}\right\|_2 \left\|\Big(\mathbf{x}^\top_{-p+1,-k}\mathbf{x}_{-p+1,-k}\Big)^{-1}\right\|_2 \left\|\mathbf{x}^\top_{-p+1,-k}\mathbf{z}_k\right\|_2$$

$$
\begin{aligned}
&= \mathcal{O}_p\Big(\sqrt{T}\Big)\mathcal{O}_p\Big(\frac{1}{T}\Big)\mathcal{O}_p\Big(\sqrt{T}\Big) = \mathcal{O}_p(1) \\
\Big\|\boldsymbol{\xi}_j^\tau - \hat{\boldsymbol{\xi}}_j^\tau\Big\|_2^2 &= \big\|P_{\mathbf{x}_{-\tau}}\boldsymbol{\xi}_j^\tau\big\|_2^2 \leq \Big\|(\boldsymbol{\xi}_j^\tau)^\top \mathbf{x}_{-\tau-1\_p}\Big\|_2 \Big\|\Big(\mathbf{x}_{-\tau-1\_p}^\top \mathbf{x}_{-\tau-1\_p}\Big)^{-1}\Big\|_2 \Big\|\mathbf{x}_{-\tau-1\_p}^\top \boldsymbol{\xi}_j^\tau\Big\|_2 \\
&= \mathcal{O}_p\Big(\sqrt{T}\Big)\mathcal{O}_p\Big(\frac{1}{T}\Big)\mathcal{O}_p\Big(\sqrt{T}\Big) = \mathcal{O}_p(1) \\
\|\mathbf{a}_j - \hat{\mathbf{a}}_j\|_2 &= \Big\|\Big(\mathbf{x}_{-\tau-1\_p}^\top \mathbf{x}_{-\tau-1\_p}\Big)^{-1}\mathbf{x}_{-\tau-1\_p}^\top \boldsymbol{\xi}_j^\tau\Big\|_2 \leq \Big\|\Big(\mathbf{x}_{-\tau-1\_p}^\top \mathbf{x}_{-\tau-1\_p}\Big)^{-1}\Big\|_2 \Big\|\mathbf{x}_{-\tau-1\_p}^\top \boldsymbol{\xi}_j^\tau\Big\|_2 \\
&= \mathcal{O}_p\Big(\frac{1}{T}\Big)\mathcal{O}_p\Big(\sqrt{T}\Big) = \mathcal{O}_p\Big(\frac{1}{\sqrt{T}}\Big) \\
\Big\|\boldsymbol{\xi}_j^\tau - \hat{\boldsymbol{\xi}}_j^\tau\Big\|_\infty &= \|\mathbf{x}_{-\tau-1\_p}(\mathbf{a}_j - \hat{\mathbf{a}}_j)\|_\infty \leq \|\mathbf{x}_{-\tau-1\_p}\|_\infty \|\mathbf{a}_j - \hat{\mathbf{a}}_j\|_1 \leq \|\mathbf{x}_{-\tau-1\_p}\|_\infty \sqrt{pd}\|\mathbf{a}_j - \hat{\mathbf{a}}_j\|_2 \\
&= o_p(1).
\end{aligned}
$$

We use $P_{-k} = \mathbf{x}_{\_p+1,-k}\Big(\mathbf{x}_{\_p+1,-k}^\top \mathbf{x}_{\_p+1,-k}\Big)^{-1}\mathbf{x}_{\_p+1,-k}^\top$ in the first equality and the according decomposition for $P_{\mathbf{x}_{-\tau}}$ in the second equality. For matrices, $\|\cdot\|_2$ denotes the spectral norm. The last equality follows as with the fourth moment assumption the maximum grows no faster than $\mathcal{O}_p\big(T^{1/4}\big)$.

Assess the numerator of the least squares coefficient

$$
\begin{aligned}
&\Big|\mathbf{z}_k^\top f\big(\boldsymbol{\xi}_j^\tau\big) - \hat{\mathbf{z}}_k^\top f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big)\Big| \\
\leq &\Big|\mathbf{z}_k^\top\Big(f\big(\boldsymbol{\xi}_j^\tau\big) - f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big)\Big)\Big| + \Big|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top f\big(\boldsymbol{\xi}_j^\tau\big)\Big| + \Big|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top\Big(f\big(\boldsymbol{\xi}_j^\tau\big) - f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big)\Big)\Big| \\
\leq &\|\mathbf{z}_k\|_2\Big\|f\big(\boldsymbol{\xi}_j^\tau\big) - f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big)\Big\|_2 + \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|_2\big\|f\big(\boldsymbol{\xi}_j^\tau\big)\big\|_2 + \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|_2\Big\|f\big(\boldsymbol{\xi}_j^\tau\big) - f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big)\Big\|_2
\end{aligned}
$$

By the moment assumption, $\|\mathbf{z}_k\|_2 = \mathcal{O}_p\Big(\sqrt{T}\Big)$ and $\big\|f\big(\boldsymbol{\xi}_j^\tau\big)\big\|_2 = \mathcal{O}_p\Big(\sqrt{T}\Big)$. We consider the difference in the nonlinearities. First, apply Taylor's theorem

$$
f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big) - f\big(\boldsymbol{\xi}_j^\tau\big) = \Big(f'\big(\boldsymbol{\xi}_j^\tau\big) + h_1\Big(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\Big)\Big) \odot \Big(\hat{\boldsymbol{\xi}}_j^\tau - \boldsymbol{\xi}_j^\tau\Big),
$$

where $h_1(\cdot)$ is the Peano form of the remainder. All functions are meant to be applied elementwise and $\odot$ denotes elementwise multiplication. For $\hat{\xi}_{t,j}^\tau \to \xi_{t,j}^\tau$, it holds $h_1\Big(\hat{\xi}_{t,j}^\tau, \xi_{t,j}^\tau\Big) \to 0$. Then,

$$
\begin{aligned}
\Big\|f\big(\boldsymbol{\xi}_j^\tau\big) - f\Big(\hat{\boldsymbol{\xi}}_j^\tau\Big)\Big\|_2^2 &= \sum_{t=1}^T\Big(f\big(\xi_{t,j}^\tau\big) - f\Big(\hat{\xi}_{t,j}^\tau\Big)\Big)^2 = \sum_{t=1}^T\Big(f'\big(\xi_{t,j}^\tau\big) + h_1\Big(\hat{\xi}_{t,j}^\tau, \xi_{t,j}^\tau\Big)\Big)^2\Big(\xi_{t,j}^\tau - \hat{\xi}_{t,j}^\tau\Big)^2 \\
&\leq \Big\|f'\big(\boldsymbol{\xi}_j^\tau\big) + h_1\Big(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\Big)\Big\|_\infty^2 \sum_{t=1}^T\Big(\xi_{t,j}^\tau - \hat{\xi}_{t,j}^\tau\Big)^2 \\
&= \Big\|f'\big(\boldsymbol{\xi}_j^\tau\big) + h_1\Big(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\Big)\Big\|_\infty^2 \Big\|\boldsymbol{\xi}_j^\tau - \hat{\boldsymbol{\xi}}_j^\tau\Big\|_2^2 = o_p(T).
\end{aligned}
$$

The maximum norm for $f'\left(\boldsymbol{\xi}_j^\tau\right)$ can be bounded at $\mathcal{O}_p\left(\sqrt{T}\right)$ by the moment assumption, and that for $h_1\left(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\right)$ is $\mathcal{O}_p(1)$ by the properties of the remainder. In summary,

$$\frac{1}{T}\hat{\mathbf{z}}_k^\top f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) = \frac{1}{T}\mathbf{z}_k^\top f\left(\boldsymbol{\xi}_j^\tau\right) + \mathcal{O}_p(1) = \mathbb{E}\left[z_{t,k} f\left(\xi_{t+\tau,j}^\tau\right)\right] + \mathcal{O}_p(1).$$

Consider the denominator

$$\begin{aligned}\left|\mathbf{z}_k^\top \mathbf{z}_k - \hat{\mathbf{z}}_k^\top \hat{\mathbf{z}}_k\right| = \left|\mathbf{z}_k^\top \mathbf{z}_k - \mathbf{z}_k^\top P_{-k}^\perp \mathbf{z}_k\right| = \left|\mathbf{z}_k^\top \left(I - P_{-k}^\perp\right)\mathbf{z}_k\right| = \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|_2^2 = \mathcal{O}_p(1) \quad \text{such that} \\ \frac{1}{T}\hat{\mathbf{z}}_k^\top \hat{\mathbf{z}}_k = \frac{1}{T}\mathbf{z}_k^\top \mathbf{z}_k + \mathcal{O}_p(1) = \mathbb{E}\left[z_{t,k}^2\right] + \mathcal{O}_p(1).\end{aligned}$$

Hence, $\hat{\beta}_k^{f,j,\tau}$ is indeed a consistent estimator.

For non-ancestors, we require a faster convergence for the numerator term. Due to orthogonality of the residuals and as $\beta_k^{f,j,\tau} = 0$

$$\hat{\mathbf{z}}_k^\top f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) = \hat{\mathbf{z}}_k^\top \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0 \boldsymbol{\beta}^{f,j,\tau}\right).$$

Assess the approximation error of this term

$$\begin{aligned}&\left|\mathbf{z}_k^\top \left(f\left(\boldsymbol{\xi}_j^\tau\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right) - \hat{\mathbf{z}}_k^\top \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right| \leq \\ &\left|\mathbf{z}_k^\top \left(\left(f\left(\boldsymbol{\xi}_j^\tau\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right) - \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right)\right| + \left|\left(\mathbf{z}_k - \hat{\mathbf{z}}_k\right)^\top \left(f\left(\boldsymbol{\xi}_j^\tau\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right| + \\ &\left|\left(\mathbf{z}_k - \hat{\mathbf{z}}_k\right)^\top \left(\left(f\left(\boldsymbol{\xi}_j^\tau\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right) - \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right)\right|.\end{aligned}$$

The last term is controlled at $\mathcal{O}_p\left(\sqrt{T}\right)$ with the Cauchy-Schwarz inequality using the rates from before. For the others, we make use of the structure of the process. Apply Taylor's theorem again

$$f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - f\left(\boldsymbol{\xi}_j^\tau\right) = \left(f'\left(\boldsymbol{\xi}_j^\tau\right) + h_1\left(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\right)\right) \odot \left(\hat{\boldsymbol{\xi}}_j^\tau - \boldsymbol{\xi}_j^\tau\right) = \left(f'\left(\boldsymbol{\xi}_j^\tau\right) + h_1\left(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\right)\right) \odot \left(\mathbf{x}_{-\tau-1\_p}(\mathbf{a}_j - \hat{\mathbf{a}}_j)\right),$$

Hence,

$$\begin{aligned}\left|\mathbf{z}_k^\top \left(f\left(\boldsymbol{\xi}_j^\tau\right) - f\left(\hat{\boldsymbol{\xi}}_j^\tau\right)\right)\right| &\leq \left\|\mathbf{z}_k^\top \left(\left(f'\left(\boldsymbol{\xi}_j^\tau\right) + h_1\left(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\right)\right) \odot \mathbf{x}_{-\tau-1\_p}\right)\right\|_2 \|\mathbf{a}_j - \hat{\mathbf{a}}_j\|_2 \\ &\leq \left(\left\|\mathbf{z}_k^\top \left(f'\left(\boldsymbol{\xi}_j^\tau\right) \odot \mathbf{x}_{-\tau-1\_p}\right)\right\|_2 + \left\|\mathbf{z}_k^\top \left(h_1\left(\hat{\boldsymbol{\xi}}_j^\tau, \boldsymbol{\xi}_j^\tau\right) \odot \mathbf{x}_{-\tau-1\_p}\right)\right\|_2\right) \|\mathbf{a}_j - \hat{\mathbf{a}}_j\|_2.\end{aligned}$$

Note that these norms are over fixed dimensional vectors such that controlling one element of the vector is the same as controlling the norm. For the first summand, use $z_{t,k} \perp \xi_{t+\tau,j}^\tau, \mathbf{x}_{t-1\_p}$ such that

the sum is over mean 0 terms and hence $\mathcal{O}_p(T)$. Consider any element in the second vector

$$\left|\sum z_{t,k} h_1\left(\hat{\xi}^{\tau}_{t+\tau,j}, \xi^{\tau}_{t+\tau,j}\right) x_{t-t',l}\right| \leq \left\|h_1\left(\hat{\boldsymbol{\xi}}^{\tau}_j, \boldsymbol{\xi}^{\tau}_j\right)\right\|_{\infty} \sum \left|z_{t,k} x_{t-t',l}\right|.$$

As we control $\left\|\boldsymbol{\xi}^{\tau}_j - \hat{\boldsymbol{\xi}}^{\tau}_j\right\|_{\infty}$ the first factor is $\mathcal{O}_p(1)$ while as the sum is $\mathcal{O}_p(T)$ such that this term is controlled as $\mathcal{O}_p(T)$ as well. The argument for

$$\left|\mathbf{z}_k^{\top}\left(\boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau} - \hat{\boldsymbol{\xi}}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right|$$

follows from the same principles without the remainder term in the Taylor expansion. As $\|\mathbf{a}_j - \hat{\mathbf{a}}_j\|_2 = \mathcal{O}_p\left(\frac{1}{\sqrt{T}}\right)$, these terms are $\mathcal{O}_p\left(\sqrt{T}\right)$. It remains to look at the middle term.

$$\begin{aligned}
&\left|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^{\top}\left(f\left(\boldsymbol{\xi}^{\tau}_j\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right| = \left|\mathbf{z}_k^{\top} P_{-k}\left(f\left(\boldsymbol{\xi}^{\tau}_j\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right| \leq \\
&\left\|\mathbf{z}_k^{\top} \mathbf{x}_{-p+1,-k}\right\|_2 \left\|\left(\mathbf{x}^{\top}_{-p+1,-k} \mathbf{x}_{-p+1,-k}\right)^{-1}\right\|_2 \left\|\mathbf{x}^{\top}_{-p+1,-k}\left(f\left(\boldsymbol{\xi}^{\tau}_j\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right)\right\|_2.
\end{aligned}$$

The first two factors are $\mathcal{O}_p\left(\sqrt{T}\right)\mathcal{O}_p\left(\frac{1}{T}\right)$ as before. In the last one, all sums are over mean 0 terms. Columns of $\mathbf{x}_{-p+1,-k}$ corresponding to time steps before t are independent of $\boldsymbol{\xi}^{\tau}_j$ and $\boldsymbol{\xi}^0$. For the other columns, orthogonality is implied as $\boldsymbol{\beta}^{f,j,\tau}$ is the least squares coefficient. Thus, this factor is $\mathcal{O}_p(T)$ and the term is $\mathcal{O}_p\left(\sqrt{T}\right)$.

$$\begin{aligned}
\frac{1}{\sqrt{T}} \hat{\mathbf{z}}_k^{\top} f\left(\hat{\boldsymbol{\xi}}^{\tau}_j\right) &= \frac{1}{\sqrt{T}} \hat{\mathbf{z}}_k^{\top}\left(f\left(\hat{\boldsymbol{\xi}}^{\tau}_j\right) - \hat{\boldsymbol{\xi}}^0 \boldsymbol{\beta}^{f,j,\tau}\right) = \frac{1}{\sqrt{T}} \mathbf{z}_k^{\top}\left(f\left(\boldsymbol{\xi}^{\tau}_j\right) - \boldsymbol{\xi}^0 \boldsymbol{\beta}^{f,j,\tau}\right) + \mathcal{O}_p(1) \\
&= \frac{1}{\sqrt{T}} \mathbf{z}_k^{\top} \mathbf{w}_k + \mathcal{O}_p(1) \xrightarrow{\mathbb{D}} \mathcal{N}\left\{0, \frac{1}{T}\mathbb{E}\left[\left(\mathbf{z}_k^{\top} \mathbf{w}_k\right)^2\right]\right\},
\end{aligned}$$

where we use the central limit theorem and Slutsky's theorem. By construction $\mathbb{E}[z_{t,k} w_{t,k}] = 0$. As $\mathbf{z}_k$, and $\mathbf{w}_k$ have time dependence only over a limited interval, the central limit theorem (Brockwell and Davis, 2009, Theorem 6.4.2) can be applied.

$$\mathbb{E}\left[\left(\mathbf{z}_k^{\top} \mathbf{w}_k\right)^2\right] = \mathbb{E}\left[\left(\sum_t z_{t,k} w_{t,k}\right)^2\right] = \sum_t \sum_{t'} \mathbb{E}\left[z_{t,k} w_{t,k} z_{t',k} w_{t',k}\right]$$

It holds $z_{t,k} \perp w_{t,k}$ and $z_{t,k} \perp z_{t',k}$ for $t \neq t'$ as there is no time-dependence. Also, $z_{t,k} \perp w_{t',k}$ for $t' > t$ as the innovation terms in $\xi^{\tau}_{t'+\tau,j}$ and $\boldsymbol{\xi}^0_{t'}$ are from later time steps. Thus, for $t' > t$

$$\mathbb{E}\left[z_{t,k} w_{t,k} z_{t',k} w_{t',k}\right] = \mathbb{E}[z_{t,k}]\mathbb{E}\left[w_{t,k} z_{t',k} w_{t',k}\right] = 0.$$

The argument for $t' < t$ is equivalent such that

$$\mathbb{E}\left[\left(\mathbf{z}_k^\top \mathbf{w}_k\right)^2\right] = \sum_t \mathbb{E}[z_{t,k}^2 w_{t,k}^2] = T\mathbb{E}[z_{t,k}^2]\mathbb{E}[w_{t,k}^2].$$

Plugging this in and using Slutsky's theorem again

$$\sqrt{T}\hat{\beta}_k^{f,j,\tau} \xrightarrow{\mathbb{D}} \mathcal{N}\left\{0, \frac{\mathbb{E}\left[w_{t,k}^2\right]}{\mathbb{E}\left[z_{t,k}^2\right]}\right\} \tag{2.8}$$

For the variance estimate, we use

$$\left\|f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right\|_2^2 = \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right)^\top \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right),$$

for which we have

$$\begin{aligned}
&\left|\left(f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right)^\top \left(f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right) - \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right)^\top \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right)\right| \leq \\
&2\left\|f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right\|_2 \left\|\left(f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right) - \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right)\right\|_2 + \\
&\left\|\left(f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right) - \left(f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right)\right\|_2^2 = \mathcal{O}_p(T) \quad \text{such that} \\
&\left\|f\left(\hat{\boldsymbol{\xi}}_j^\tau\right) - \hat{\boldsymbol{\xi}}^0\hat{\boldsymbol{\beta}}^{f,j,\tau}\right\|_2^2/T = \left\|f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right\|_2^2/T + \mathcal{O}_p(1) = \mathbb{E}[w_{t,k}^2] + \mathcal{O}_p(1) = \mathcal{O}_p(1).
\end{aligned}$$

The law of large numbers applies here as

$$\left\|f(\boldsymbol{\xi}_j^\tau) - \boldsymbol{\xi}^0\boldsymbol{\beta}^{f,j,\tau}\right\|_2^2/T$$

can be split into several converging sums of i.i.d. random variables. Thus, we get

$$\hat{\sigma}^2 = \mathcal{O}_p(1) \quad \text{and} \quad \widehat{\text{var}}\left(\hat{\beta}_l^{f,j,\tau}\right) = \mathcal{O}_p\left(\frac{1}{T}\right) \quad \forall l.$$

For non-ancestors k,

$$\widehat{\text{var}}\left(\sqrt{T}\hat{\beta}_l^{f,j,\tau}\right) = T\widehat{\text{var}}\left(\hat{\beta}_l^{f,j,\tau}\right) \xrightarrow{\mathbb{P}} \frac{\mathbb{E}\left[w_{t,k}^2\right]}{\mathbb{E}\left[z_{t,k}^2\right]},$$

i.e., the estimated variance approaches the asymptotic variance leading to the desired standard normal pivot.

2.A.5 Near-epoch dependence

We adapt the concept of near-epoch dependence, see, e.g., Davidson and de Jong (1997); Davidson (2002). Define for $\tau \leq t$ $\mathfrak{F}_\tau^t = \sigma(\boldsymbol{\epsilon}_\tau, \ldots, \boldsymbol{\epsilon}_t)$ the σ-field generated by a subset of the innovation terms and $\mathbb{E}_{t-m}^{t+m}[\cdot]$ the conditional expectation given $\mathfrak{F}_{t-m}^{t+m}$. Let y_t be some random process.

Definition 2.1. y_t is near-epoch dependent on $\{\boldsymbol{\epsilon}_t\}$ in L_p-norm, say L_p-NED, for $p > 0$ if

$$\mathbb{E}\big[\big(y_t - \mathbb{E}_{t-m}^{t+m}[y_t]\big)^p\big]^{1/p} \leq d_t \nu(m)$$

where d_t is a sequence of positive constants, and $\nu(m) \overset{m\to\infty}{\to} 0$. It is said to be L_p-NED of size $-\mu$ if $\nu(m) = \mathcal{O}\big(m^{-\mu-\delta}\big)$ for some $\delta > 0$. It is said to be geometrically L_p-NED if $\nu(m) = \mathcal{O}(\exp(-\delta m))$ for some $\delta > 0$.

By establishing near-epoch dependence, we can apply the law of large numbers and the central limit theorem in appropriate places.

Lemma 2.1. Let $\mathbf{x}_t$ follow an SVAR (2.1) with finite second moments for which (A2.1) - (A2.3) holds. Then, $x_{t,k}$ is geometrically L_2-NED on $\{\boldsymbol{\epsilon}_t\}$ for every $k \in \{1, \ldots\}$.

By the triangle inequality, the same holds then for every finite linear combination of $\mathbf{x}_t$.

Lemma 2.2. Let $\mathbf{x}_t$ follow an SVAR (2.1) with finite second moments for which (A2.1) - (A2.3) holds. Then, $x_{t,k}x_{t+\tau,l}$ is geometrically L_1-NED on $\{\boldsymbol{\epsilon}_t\}$ for every $l \in \{1, \ldots, d\}$ and $\tau < \infty$.

Lemma 2.3. Let $\mathbf{x}_t$ follow an SVAR (2.1) with finite second moments for which (A2.1) - (A2.3) holds. Let ξ_t be a quantity determined by $\mathfrak{F}_{t-\tau}^{t+\tau}$ for some $\tau < \infty$ such that $\mathbb{E}[\xi_t]$ and $\mathbb{E}[x_{t,k}\xi_t]$ are finite. Then, $x_{t,k}\xi_t$ is geometrically L_1-NED on $\{\boldsymbol{\epsilon}_t\}$.

As the innovation sequence $\{\boldsymbol{\epsilon}_t\}$ is independent over time, it satisfies every mixing property. Thus, we can apply Davidson and de Jong (1997)[Theorem 3.3] to obtain the LLN for L_1-NED quantities

2.A.5.1 Proof of Lemma 2.1

Consider the form

$$\mathbf{x}_{t\_p} = \tilde{\mathbf{B}}\mathbf{x}_{t-1\_p} + \boldsymbol{\xi}_{t\_p} = \sum_{\tau=0}^{\infty} \tilde{\mathbf{B}}^\tau \boldsymbol{\xi}_{t-\tau\_p}$$

as in Lütkepohl (2005) and (2.3). Then

$$\mathbf{x}_{t\_p} - \mathbb{E}_{t-m}^{t+m}[\mathbf{x}_{t\_p}] = \mathbf{x}_{t\_p} - \mathbb{E}_{t-m}^{t}[\mathbf{x}_{t\_p}] = \sum_{\tau=m+1}^{\infty} \tilde{\mathbf{B}}^\tau \boldsymbol{\xi}_{t-\tau\_p}.$$

Consider any $x_{t,k} = \mathbf{e}_k^\top \mathbf{x}_{t\_p}$, where $\mathbf{e}_k \in \mathbb{R}^{dp}$ is the according unit vector.

$$\begin{aligned}
\mathbb{E}\Big[\big(x_{t,k} - \mathbb{E}^t_{t-m}[x_{t,k}]\big)^2\Big] &= \mathbf{e}_k^\top \mathbb{E}\left[\sum_{\tau=m+1}^{\infty} \tilde{\mathbf{B}}^\tau \boldsymbol{\xi}_{t-\tau_p} \left(\sum_{\tau'=m+1}^{\infty} \tilde{\mathbf{B}}^{\tau'} \boldsymbol{\xi}_{t-\tau'_p}\right)^\top\right] \mathbf{e}_k \\
&= \mathbf{e}_k^\top \sum_{\tau=m+1}^{\infty} \tilde{\mathbf{B}}^\tau \mathbb{E}\Big[\boldsymbol{\xi}_{t_p} \boldsymbol{\xi}_{t_p}^\top\Big] \Big(\tilde{\mathbf{B}}^\tau\Big)^\top \mathbf{e}_k \\
&\leq \sum_{\tau=m+1}^{\infty} \Big\|\mathbf{e}_k^\top \tilde{\mathbf{B}}^\tau\Big\|_2^2 \lambda_{\max}\Big(\mathbb{E}\Big[\boldsymbol{\xi}_{t_p} \boldsymbol{\xi}_{t_p}^\top\Big]\Big) \leq C \sum_{\tau=m+1}^{\infty} \lambda_{\max}\Big(\tilde{\mathbf{B}}\Big)^{2\tau},
\end{aligned}$$

where $\lambda_{\max}(\cdot)$ denotes the largest absolute eigenvalue of a matrix, and C is a accordingly chosen constant. Under the stability assumption (A2.3), $\lambda_{\max}\Big(\tilde{\mathbf{B}}\Big) < 1$ such that the sum decreases as

$$\mathcal{O}\Big(\lambda_{\max}\Big(\tilde{\mathbf{B}}\Big)\Big)^{2m} \coloneqq \mathcal{O}(\exp(-2\theta m)), \quad \text{where} \quad \theta = -\log\Big(\lambda_{\max}\Big(\tilde{\mathbf{B}}\Big)\Big) > 0.$$

2.A.5.2 Proof of Lemma 2.2

As $x_{t,k}$ is a linear combination of innovation terms, we can write

$$\begin{aligned}
x_{t,k} &= \mathbb{E}^{t+\tau}_{t-m}[x_{t,k}] + \mathbb{E}^{t-m-1}_{-\infty}[x_{t,k}] \coloneqq \tilde{x}_{t,k} + \hat{x}_{t,k} \quad \text{and analogously} \\
x_{t+\tau,l} &= \mathbb{E}^{t+\tau}_{t-m}[x_{t+\tau,l}] + \mathbb{E}^{t-m-1}_{-\infty}[x_{t+\tau,l}] \coloneqq \tilde{x}_{t+\tau,l} + \hat{x}_{t+\tau,l}.
\end{aligned}$$

Let $m \geq \tau$.

$$\begin{aligned}
&\mathbb{E}\big[\big|x_{t,k}x_{t+\tau,l} - \mathbb{E}^{t+m}_{t-m}[x_{t,k}x_{t+\tau,l}]\big|\big] = \mathbb{E}\big[\big|x_{t,k}x_{t+\tau,l} - \mathbb{E}^{t+\tau}_{t-m}[x_{t,k}x_{t+\tau,l}]\big|\big] \\
=&\mathbb{E}\big[\big|(\tilde{x}_{t,k} + \hat{x}_{t,k})(\tilde{x}_{t+\tau,l} + \hat{x}_{t+\tau,l}) - \mathbb{E}^{t+\tau}_{t-m}[(\tilde{x}_{t,k} + \hat{x}_{t,k})(\tilde{x}_{t+\tau,l} + \hat{x}_{t+\tau,l})]\big|\big] \\
=&\mathbb{E}[|\tilde{x}_{t,k}(\hat{x}_{t+\tau,l} - \mathbb{E}[\hat{x}_{t+\tau,l}]) + \tilde{x}_{t+\tau,l}(\hat{x}_{t,k} - \mathbb{E}[\hat{x}_{t,k}]) + \hat{x}_{t,k}\hat{x}_{t+\tau,l} - \mathbb{E}[\hat{x}_{t,k}\hat{x}_{t+\tau,l}]|] \\
\leq&2(\mathbb{E}[|\tilde{x}_{t,k}|]\mathbb{E}[|\hat{x}_{t+\tau,l}|] + \mathbb{E}[|\tilde{x}_{t+\tau,l}|]\mathbb{E}[|\hat{x}_{t,k}|] + \mathbb{E}[|\hat{x}_{t,k}\hat{x}_{t+\tau,l}|]).
\end{aligned}$$

As argued before $\hat{x}_{t,k}, \hat{x}_{t+\tau,l}$ have exponentially decreasing moments while as the moments of $\tilde{x}_{t,k}$, $\tilde{x}_{t+\tau,l}$ are bounded by the assumptions of the process. So, overall this sum is exponentially decreasing which establishes the Lemma.

2.A.5.3 Proof of Lemma 2.3

Decompose

$$x_{t,k} = \tilde{x}_{t,k} + \hat{x}_{t,k}$$

as before. Let $m \geq \tau$.

$$\begin{aligned}&\mathbb{E}\big[\big|x_{t,k}\xi_t - \mathbb{E}_{t-m}^{t+m}[x_{t,k}\xi_t]\big|\big] = \mathbb{E}\big[\big|(\tilde{x}_{t,k}+\hat{x}_{t,k})\xi_t - \mathbb{E}_{t-m}^{t+m}[(\tilde{x}_{t,k}+\hat{x}_{t,k})\xi_t]\big|\big]\\ =&\mathbb{E}\big[\big|\xi_t\big(\hat{x}_{t,k} - \mathbb{E}_{t-m}^{t+m}[\hat{x}_{t,k}]\big)\big|\big] \leq 2\mathbb{E}[|\xi_t|]\mathbb{E}[|\hat{x}_{t,k}|],\end{aligned}$$

which attains the exponential rate as argued before.

2.A.6 Proof of Theorem 2.2

For simplicity, assume $\mathrm{MA}^{\tau\to j}(k) = \mathrm{MA}(k)$ such that

$$z_{t,k} = x_{t,k} - \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}^\top \boldsymbol{\gamma}^{\tau\to j,k}.$$

We have

$$\begin{aligned}\beta_k^{f,j,\tau} = 0\ \forall f(\cdot) &\iff \mathbb{E}\big[z_{t,k} f\big(\xi_{t+\tau,j}^\tau\big)\big] = \mathbb{E}\big[\mathbb{E}\big[z_{t,k} \mid \xi_{t+\tau,j}^\tau\big] f\big(\xi_{t+\tau,j}^\tau\big)\big] = 0\ \forall f(\cdot)\\ &\iff \mathbb{E}\big[z_{t,k} \mid \xi_{t+\tau,j}^\tau\big] = 0.\end{aligned}$$

Using the equality above, the last condition is equivalent to the one in the theorem. By construction

$$x_{t+\tau,j} = x_{t,k}\alpha + \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}^\top \boldsymbol{\beta} + \tilde{\epsilon} \quad \text{for some } \alpha, \boldsymbol{\beta},$$

where $\tilde{\epsilon}$ is a linear combination of $x_{t',l}$ whose paths to $x_{t,k}$ are blocked by $\mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}$ and noise terms $\epsilon_{t',l}$ with $t' > t$ such that $x_{t,k} \perp \tilde{\epsilon} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}$. Hence,

$$x_{t,k} \not\perp x_{t+\tau,j} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)} \implies \alpha \neq 0.$$

Then,

$$\mathbb{E}\big[z_{t,k}\xi_{t+\tau,j}^\tau\big] = \mathbb{E}[z_{t,k}x_{t+\tau,j}] = \alpha\mathbb{E}[z_{t,k}x_{t,k}] + \mathbb{E}\Big[z_{t,k}\mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}^\top\Big]\boldsymbol{\beta} + \mathbb{E}[z_{t,k}\tilde{\epsilon}] = \alpha\mathbb{E}\big[z_{t,k}^2\big] \neq 0,$$

i.e., the identity function leads to a non-zero regression coefficient if the conditional independence does not hold. The first equality holds as $z_{t,k}$ is independent from $\mathbf{x}_{t'}$ for $t' < t$. In the next expression, the middle summand vanishes by the construction of the regression residual $z_{t,k}$. For the last summand, note that all contributions from $\epsilon_{t',l}$ with $t' > t$ are independent of $z_{t,k}$ trivially, while as contributions of other $x_{t',l}$ must be uncorrelated from $z_{t,k}$ as it is a least squares residual.

For the last part, assume $x_{t,k} \perp x_{t+\tau,j} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}$. This is equivalent to $x_{t,k} \perp \xi_{t+\tau,j}^\tau \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}$ as $\xi_{t+\tau,j}^\tau - x_{t+\tau,j}$ only depends on time before t such that $x_{t,k}$ cannot depend

on it given $\mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}$. Then, it follows

$$\begin{aligned}\mathbb{E}\big[z_{t,k} \mid \xi^{\tau}_{t+\tau,j}\big] &= \mathbb{E}\Big[x_{t,k} - \mathbf{x}^{\top}_{t,\mathrm{MA}^{\tau\to j}(k)}\boldsymbol{\gamma}^{\tau\to j,k} \mid \xi^{\tau}_{t+\tau,j}\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[x_{t,k} - \mathbf{x}^{\top}_{t,\mathrm{MA}^{\tau\to j}(k)}\boldsymbol{\gamma}^{\tau\to j,k} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}, \xi^{\tau}_{t+\tau,j}\Big] \mid \xi^{\tau}_{t+\tau,j}\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big[x_{t,k} \mid \mathbf{x}_{t,\mathrm{MA}^{\tau\to j}(k)}\Big] - \mathbf{x}^{\top}_{t,\mathrm{MA}^{\tau\to j}(k)}\boldsymbol{\gamma}^{\tau\to j,k} \mid \xi^{\tau}_{t+\tau,j}\Big] = 0,\end{aligned}$$

where the second to last equality uses conditional independence and the last follows trivially from the assumption.

Finally, we can consider the general case where $\mathrm{MA}^{\tau\to j}(k) \subseteq \mathrm{MA}(k)$. Let $l \in \mathrm{CH}^0(k) \setminus \mathrm{AN}^{\tau}(j)$. We know that $\xi_{t,l}$ has a regression coefficient of 0 by Theorem 2.1. Hence, removing it from the model cannot change the remaining least squares parameters. Now, after removing $x_{t,l}$, its parents do not contribute to $z_{t,k}$ unless they are in $\mathrm{MA}^{\tau\to j}(k)$ due to one of the other conditions. Thus, removing these additionally cannot change $z_{t,k}$ and hence $\beta^{f,j,\tau}_k$. Therefore, it suffices to analyze with $\mathrm{MA}^{\tau\to j}(k)$ only.

2.A.7 Combined p-values

The arguments presented here follow mainly on Meinshausen et al. (2009). But, by noting that one should focus on the order statistics and not continuous quantiles, we slightly improve the penalty term for the combined p-value. Also, we omit here the possibility of ignoring the lowest p-values as there might be cases where only one should be non-uniform.

Let $x_{t,j}$ and $x_{t,k}$ be two of the observed time series and $p^{j,0}_k, \dots p^{j,p}_k$ all p-values for potential effects from k to j. Sort these p-values from lowest to largest, say $p^{j}_{k,(1)}, \dots p^{j}_{k,(r)}$ where $r = p + 1$. We get our combined p-value as

$$p^{j}_k = \min_{i\in\{1,\dots,r\}} \frac{r}{i} p^{j}_{k,(i)} \sum_{i'=1}^{r} \frac{1}{i'}.$$

Proposition 2.1. If $p^{j,0}_k, \dots p^{j,p}_k$ are all Uniform$(0,1)$, p^{j}_k is a valid p-value, i.e.,

$$\mathbb{P}\Big(p^{j}_k \le \alpha\Big) \le \alpha \ \forall \alpha \in (0,1).$$

Proof

Define

$$\pi^{j}_k(u) = \frac{1}{r}\sum_{i=1}^{r} 1\Big\{p^{j}_{k,(i)} \le u\Big\}.$$

We have

$$\pi^{j}_k\Big(\alpha\frac{i}{r}\Big) \ge \frac{i}{r} \iff p^{j}_{k,(i)}\frac{r}{i} \le \alpha$$

Let U be a random variable in $[0,1]$ and consider

$$\max_{i\in\{1,\dots,r\}} \frac{1\{U \le \alpha i/r\}}{i/r} = \begin{cases} 0 & U > \alpha \\ \dfrac{r}{\lceil Ur/\alpha \rceil} & \text{otherwise.} \end{cases}$$

If U is uniformly distributed we get the expectation

$$\mathbb{E}\left[\max_{i\in\{1,\dots,r\}} \frac{1\{U \le \alpha i/r\}}{i/r}\right] = \alpha \sum_{i=1}^{r} \frac{1}{i}$$

as for each possible i, there is a segment of length α/r where $\lceil Ur/\alpha \rceil = i$.

Thus, if the individual p-values are uniform

$$\mathbb{E}\left[\max_{i\in\{1,\dots,r\}} \frac{1}{r}\sum_{\tau=0}^{p} \frac{1\left\{p_k^{j,\tau} \le \alpha i/r\right\}}{i/r}\right] \le \frac{1}{r}\sum_{\tau=0}^{p} \mathbb{E}\left[\max_{i\in\{1,\dots,r\}} \frac{1\left\{p_k^{j,\tau} \le \alpha i/r\right\}}{i/r}\right] = \alpha \sum_{i=1}^{r} \frac{1}{i}.$$

Apply the definition of $\pi_k^j(\cdot)$ and the Markov inequality.

$$\begin{aligned}
\alpha \sum_{i=1}^{r} \frac{1}{i} &\ge \mathbb{E}\left[\max_{i\in\{1,\dots,r\}} \frac{1}{r}\sum_{\tau=0}^{p} \frac{1\left\{p_k^{j,\tau} \le \alpha i/r\right\}}{i/r}\right] = \mathbb{E}\left[\max_{i\in\{1,\dots,r\}} \frac{\pi_k^j(\alpha i/r)}{i/r}\right] \ge \mathbb{P}\left(\max_{i\in\{1,\dots,r\}} \frac{\pi_k^j(\alpha i/r)}{i/r} \ge 1\right) \\
&= \mathbb{P}\left(\max_{i\in\{1,\dots,r\}} 1\left\{\pi_k^j(\alpha i/r) \ge i/r\right\} \ge 1\right) = \mathbb{P}\Big(\exists i \in \{1,\dots,r\} : \quad \pi_k^j(\alpha i/r) \ge i/r\Big) \\
&= \mathbb{P}\Big(\exists i \in \{1,\dots,r\} : p_{k,(i)}^j \frac{r}{i} \le \alpha\Big) = \mathbb{P}\left(\min_{i\in\{1,\dots,r\}} p_{k,(i)}^j \frac{r}{i} \le \alpha\right) = \mathbb{P}\left(p_k^j \le \alpha \sum_{i=1}^{r} \frac{1}{i}\right).
\end{aligned}$$

As α is arbitrary in this argument, this establishes the super-uniformity of the p-value. The only place where uniformity of the individual $p_k^{j,\tau}$ is invoked is when calculating the expectation of a bounded function. Hence, if the individual p-values are asymptotically uniform, we obtain an asymptotic result for p_k^j.

2.B Details on the simulation setup

We use the following distributions for the $\epsilon_{t,j}$: two t_7 distributions, two centered uniform distributions, a centered Laplace distribution with scale 1, and a standard normal distribution. All distributions are normalized to have unit variance. For each simulation run, we randomly permute the distributions to assign them to $\epsilon_{t,1}$ to $\epsilon_{t,6}$.

The edges $x_{t,k} \to x_{t,j}$ with $k < j$ are present, i.e., the entry $(\mathbf{B}_0)_{jk}$ is non-zero, with probability 0.2 each such that an average of 3 parental connections exists.

We assign preliminary edge weights uniformly in $[0.5, 1]$. These are further scaled such that for every $x_{t,j}$ which has instantaneous ancestors, the standard deviation of

$$\xi_{t,k} - \epsilon_{t,k}$$

is uniformly chosen from $[\sqrt{0.5}, \sqrt{2}]$ to control the signal-to-noise ratio.

To initialize the graph and the weights, we use the function `randomDAG` from the `R`-package `pcalg` (Kalisch et al., 2012) before applying our changes to the weights to enforce the constraints.

The entries in $\mathbf{B}_1$ are non-zero with probability 0.1. If so, they are sampled uniformly with absolute value in $[0.2, 0.8]$ and assigned a random sign with equal probabilities. If the maximum absolute eigenvalue of $\tilde{\mathbf{B}}$ would be larger than 0.95, $\mathbf{B}_1$ is shrunken such that this absolute eigenvalue is 0.95 to ensure stability.

We initiate the time series randomly and discard the first 10^4 observations to ensure strict stationarity (approximately).

Chapter 3

Higher-order least squares: assessing partial goodness of fit of linear causal models

Christoph Schultheiss, Peter Bühlmann, and Ming Yuan
Journal of the American Statistical Association 119 (546), 1019-1031.

Abstract

We introduce a simple diagnostic test for assessing the overall or partial goodness of fit of a linear causal model with errors being independent of the covariates. In particular, we consider situations where hidden confounding is potentially present. We develop a method and discuss its capability to distinguish between covariates that are confounded with the response by latent variables and those that are not. Thus, we provide a test and methodology for *partial* goodness of fit. The test is based on comparing a novel higher-order least squares principle with ordinary least squares. In spite of its simplicity, the proposed method is extremely general and is also proven to be valid for high-dimensional settings.

3.1 Introduction

Linear models are the most commonly used statistical tools to study the relationship between a response and a set of covariates. The regression coefficient corresponding to a particular covariate is usually interpreted as its net effect on the response variable when all else is held fixed. Such an interpretation is essential in many applications and yet could be rather misleading when the linear model assumptions are in question, in particular, when there are hidden confounders.

In this work, we develop a simple but powerful approach to goodness of fit tests for potentially high-dimensional linear causal models, including also tests for partial goodness of fit of single predictor variables. While hidden confounding is the primary alternative in mind, different nonlinear deviations from the linear model assumption are also in scope. Tests for goodness of fit tests are essential to statistical modeling (e.g., Lehmann et al., 2005) and the concept is also very popular in econometrics where it is referred to as specification tests. For an overview of such methods, see, e.g., Godfrey (1989) or Maddala and Lahiri (2009).

Another set of related works is Buja et al. (2019a,b), which elaborately discusses deviations from the (linear) model and how distributional robustness, i.e., robustenss against shifts in the covariates' distribution, links to correctly specified models. For this, they introduce the definition of "well-specified" statistical functionals. Distributional robustness, implied by well-specification, is also related to the causal interpretation of a linear model as discussed in Peters et al. (2016).

We consider here the question when and which causal effects can be inferred from the ordinary least squares estimator or a debiased Lasso procedure for the high-dimensional setting, even when there is hidden confounding. We address this by partial goodness of fit testing: if the data speaks against a linear causal model, we are able to specify which components of the least squares estimator should be rejected to be linear causal effects and which not. In the case of a joint Gaussian distribution, one cannot detect anything: this corresponds to a well-known unidentifiability result in causality (Hyvärinen and Oja, 2000; Peters et al., 2014). But, in certain models, we are able to identify some causal relations. Of particular importance are non-Gaussian linear structural equation models, as used in Shimizu et al. (2006) or Wang and Drton (2020) amongst others. The latter constructs the causal graph from observational data in a stepwise procedure using a test statistic similar to the one we suggest.

Our strategy has a very different focus than other approaches which do not rely on the least squares principle any longer to deal with the issue of hidden confounding. Most prominent, particularly in econometrics, is the framework of instrumental variables regression: assuming valid instruments, one can identify all causal effects, see, e.g., Angrist et al. (1996) or the books by Bowden and Turkington (1985) and Imbens and Rubin (2015). The popular Durbin-Wu-Hausman test (Hausman, 1978) for validity of instruments bears a relation to our methodology, namely that we are also looking at the

difference of two estimators to test goodness of fit.

Our automated partial goodness of fit methodology is easy to be used as it is based on ordinary (or high-dimensional adaptions of) least squares and its novel higher-order version: we believe that this simplicity is attractive for statistical practice.

3.1.1 Our contribution

We propose a novel method with a corresponding test, called higher-order least squares (HOLS). The test statistic is based on the residuals from an ordinary least squares or Lasso fit. In that regard, it is related to Shah and Bühlmann (2018) who use "residual prediction" to test for deviation from the linear model. However, our approach does neither assume Gaussian errors nor does it rely on sample splitting, and our novel test statistic has a $\sqrt{n}$ convergence rate (with n denoting the sample size).

In addition to presenting a "global" goodness of fit test for the entire model, we also develop a local interpretation that allows detecting which among the covariates are giving evidence for hidden confounding or nonlinear relations. Thus, we strongly increase the amount of extracted information compared to a global goodness of fit test. In particular, in the case of localized (partial) confounding in linear structural equation models, we are able to recover the unconfounded regression parameters for a subset of predictors. This is a setting where techniques assuming dense (essentially global) confounding, as in Ćevid et al. (2020) or Guo et al. (2022), fail.

The work by Buja et al. (2019a,b), especially the second paper, shows how to detect deviations in a linear model using reweighting of the data. Our HOLS technique can be seen as a special way of reweighting. In contrast to their work, we provide a simple test statistic that tests for well-specification without requiring any resampling. Furthermore, we provide guarantees for a local interpretation under suitable modeling assumptions while as their per-covariate view remains rather exploratory.

3.1.2 Outline

The remainder of this paper will be structured as follows. We conclude this section with the necessary notation. In Section 3.2, we present the main idea of HOLS and the according global null hypothesis. For illustrative purposes, we first discuss univariate regression. Then, we consider multivariate regression and extend our theory to high-dimensional problems incorporating the (debiased) Lasso. In Section 3.3, we present the local interpretation when the global null does not hold true alongside with theoretical guarantees. Models for which this local interpretation is most suitable are discussed in Section 3.4. Section 3.5 contains a real data analysis. We conclude with a summarizing discussion in Section 3.6.

3.1.3 Notation

We present some notation that is used throughout this work. Vectors and matrices are written in boldface, while scalars have the usual lettering. This holds for both random and fixed quantities. We

use upper case letters to denote a random variable, e.g., $\mathbf{X}$ or Y. We use lower case letters to denote i.i.d. copies of a random variable, e.g., $\mathbf{x}$. If $\mathbf{X} \in \mathbb{R}^p$, then $\mathbf{x} \in \mathbb{R}^{n \times p}$. With a slight abuse of notation, $\mathbf{x}$ can either denote the copies or realizations thereof. We write $\mathbf{x}_j$ to denote the j-th column of matrix $\mathbf{x}$ and $\mathbf{x}_{-j}$ to denote all but the j-th column. We write $\overset{H_0}{=}$ to state that equality holds under H_0. With $\leftarrow$, we emphasize that an equality between random variables is induced by a causal mechanism. We use $\odot$ to denote elementwise multiplication of two vectors, e.g., $\mathbf{x} \odot \mathbf{y}$. Similarly, potencies of vectors are also to be understood in an elementwise fashion, e.g., $\mathbf{x}^2 = \mathbf{x} \odot \mathbf{x}$. $\boldsymbol{I}_n$ is the n-dimensional identity matrix. $\boldsymbol{P}_{-j}$ denotes the orthogonal projection onto $\mathbf{x}_{-j}$ and $\boldsymbol{P}_{-j}^{\perp} = \boldsymbol{I}_n - \boldsymbol{P}_{-j}$ denotes the orthogonal projection onto its complement. For some random vector $\mathbf{X}$, we have the moment matrix $\boldsymbol{\Sigma}^{\mathbf{X}} \coloneqq \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big]$. Note that this equals the covariance matrix for centered $\mathbf{X}$. We denote statistical independence by $\perp$. We write $\mathbf{e}$ to denote a vector for which every entry is 1 and $\mathbf{e}_j$ to denote the unit vector in the direction of the j-th coordinate axis.

3.2 Higher-order least squares (HOLS)

We develop here the main idea of higher-order least squares (HOLS) estimation.

3.2.1 Univariate regression as a motivating case

It is instructive to begin with the case of simple linear regression where we have a pair of random variables X and Y. We consider the causal linear model

$$Y \leftarrow X\beta + \mathcal{E}, \quad \text{where} \quad X \perp \mathcal{E}, \quad \mathbb{E}[\mathcal{E}] = 0 \quad \text{and} \quad \mathbb{E}\big[\mathcal{E}^2\big] = \sigma^2 < \infty. \tag{3.1}$$

We formulate a null hypothesis that the model in (3.1) is correct and we denote such a hypothesis by H_0. This model is of interest as β describes the effect of a unit change if we were to intervene on covariate X without intervening on the independent $\mathcal{E}$. Such model, or its multivariate extension, is often assumed in causal discovery, see, e.g., Shimizu et al. (2006) or Hoyer et al. (2008b). Therefore, we aim to provide a test for its well-specification.

Estimation of the regression parameter is typically done by the least squares principle

$$\beta^{OLS} \coloneqq \underset{b \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\Big[(Y - Xb)^2\Big] = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]} \overset{H_0}{=} \beta,$$

where we use the superscript OLS to denote ordinary least squares. Alternatively, we can pre-multiply the linear model (3.1) with X: the parameter minimizing the expected squared error loss is then

$$\beta^{HOLS} \coloneqq \underset{b \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\Big[\big(XY - X^2 b\big)^2\Big] = \frac{\mathbb{E}\big[X^3 Y\big]}{\mathbb{E}[X^4]} \overset{H_0}{=} \frac{\mathbb{E}\big[X^4 \beta\big]}{\mathbb{E}[X^4]} = \beta.$$

More generally, $\beta^{HOLS} = \beta^{OLS} = \beta$, if $\mathbb{E}[Y|X] = X\beta$. Using the definition from Buja et al. (2019b), this means that the OLS parameter is well-specified. The estimation principle is called higher-order least squares, or HOLS for short, as it involves higher-order moments of X. One could also multiply the linear model with a factor other than X, which may have implications on the power to detect deviations from (3.1). We shall focus here on the specific choice to fix ideas.

The motivation to look at HOLS is when H_0 is violated, in terms of a hidden confounding variable: let H be a hidden confounder leading to a model

$$X \leftarrow \mathcal{E}_X + H\rho, \quad Y \leftarrow X\beta + H\alpha + \mathcal{E},$$

where $\mathcal{E}_X$, H, and $\mathcal{E}$ are all independent and α and ρ define additional model parameters. In particular, we can compute under such a confounding model that

$$\beta^{HOLS} - \beta^{OLS} = \rho\alpha\left(\frac{3\mathbb{E}\big[\mathcal{E}_X^2\big]\mathbb{E}\big[H^2\big] + \rho^2\mathbb{E}\big[H^4\big]}{\mathbb{E}\big[\mathcal{E}_X^4\big] + 6\rho^2\mathbb{E}\big[\mathcal{E}_X^2\big]\mathbb{E}[H^2] + \rho^4\mathbb{E}[H^4]} - \frac{\mathbb{E}\big[H^2\big]}{\mathbb{E}\big[\mathcal{E}_X^2\big] + \rho^2\mathbb{E}[H^2]}\right). \tag{3.2}$$

For simplicity, we assumed here $\mathbb{E}[\mathcal{E}_X] = \mathbb{E}[H] = \mathbb{E}[\mathcal{E}] = 0$. In practice, one can get rid of this assumption by including an intercept in the model. If either α or ρ equals to 0, we see that the difference in (3.2) is 0. This is not surprising as there is no confounding effect when either X or Y is unaffected. However, this is not the only possibility how the difference can be 0. Namely,

$$\mathbb{E}\big[H^2\big]\Big(\mathbb{E}\big[\mathcal{E}_X^4\big] - 3\mathbb{E}\big[\mathcal{E}_X^2\big]^2\Big) = \rho^2\mathbb{E}\big[\mathcal{E}_X^2\big]\Big(\mathbb{E}\big[H^4\big] - 3\mathbb{E}\big[H^2\big]^2\Big) \Rightarrow \beta^{HOLS} - \beta^{OLS} = 0.$$

Especially, if neither $\mathcal{E}_X$ nor H have excess kurtosis, the difference is 0 for any ρ. This can be intuitively explained as it corresponds to Gaussian data (up to the moments we consider). For Gaussian $\mathcal{E}_X$ and H, one can always write

$$Y = X\beta^{OLS} + \tilde{\mathcal{E}} \quad \text{where} \quad X \perp \tilde{\mathcal{E}},$$

which cannot be distinguished from the null model (3.1). Or in other words $\mathbb{E}[Y|X] = X\beta^{OLS}$, i.e., the OLS parameter is well-specified although it is not the parameter β that we would like to recover. For other data generating distributions, one should be able to distinguish H_0 from certain deviations when hidden confounding is present. We discuss the implications of this in the general multivariate setting in Section 3.3.2. Similar behaviour occurs for a violation of H_0 in terms of a nonlinear model $Y = f(X, \epsilon)$ which then (typically) leads to $\beta^{HOLS} - \beta^{OLS} \neq 0$.

One can construct a test based on the sample estimates of β^{HOLS} and β^{OLS}. We consider the centered data

$$\tilde{\mathbf{x}} = \mathbf{x} - \bar{x}\mathbf{e}, \quad \tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{e} \quad \text{and} \quad \tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} - \bar{\epsilon}\mathbf{e} = \left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^\top\right)\boldsymbol{\epsilon},$$

where we use the upper bar to denote sample means. We can derive

$$\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{e} = \mathbf{x}\beta - \bar{x}\mathbf{e}\beta + \boldsymbol{\epsilon} - \bar{\epsilon}\mathbf{e} = \tilde{\mathbf{x}}\beta + \tilde{\boldsymbol{\epsilon}}.$$

We now obtain $\hat{\beta}^{OLS}$ from regression through the origin of $\tilde{\mathbf{y}}$ versus $\tilde{\mathbf{x}}$ with an error term of $\tilde{\boldsymbol{\epsilon}}$ and $\hat{\beta}^{HOLS}$ from regression through the origin of $\tilde{\mathbf{x}} \odot \tilde{\mathbf{y}}$ versus $\tilde{\mathbf{x}}^2$ with an error term of $\tilde{\mathbf{x}} \odot \tilde{\boldsymbol{\epsilon}}$. More precisely, we define

$$\hat{\beta}^{OLS} \coloneqq \frac{\tilde{\mathbf{x}}^\top \tilde{\mathbf{y}}}{\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}} \quad \text{and} \quad \hat{\beta}^{HOLS} \coloneqq \frac{\left(\tilde{\mathbf{x}}^2\right)^\top (\tilde{\mathbf{x}} \odot \tilde{\mathbf{y}})}{(\tilde{\mathbf{x}}^2)^\top (\tilde{\mathbf{x}}^2)} = \frac{\left(\tilde{\mathbf{x}}^3\right)^\top (\tilde{\mathbf{y}})}{(\tilde{\mathbf{x}}^2)^\top (\tilde{\mathbf{x}}^2)}.$$

Under H_0, one can see that $\left(\hat{\beta}^{HOLS} - \hat{\beta}^{OLS}\right)$ given $\mathbf{x}$ is some known linear combination of $\boldsymbol{\epsilon}$. Assuming further Gaussianity of $\boldsymbol{\epsilon}$, it is conditionally Gaussian. We find

$$\left(\hat{\beta}^{HOLS} - \hat{\beta}^{OLS}\right)\Big|\mathbf{x} \overset{H_0}{\sim} \mathcal{N}\left(0, \sigma^2 \left(\frac{\left(\tilde{\mathbf{x}}^3\right)^\top \left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^\top\right)\left(\tilde{\mathbf{x}}^3\right)}{\left((\tilde{\mathbf{x}}^2)^\top (\tilde{\mathbf{x}}^2)\right)^2} - \frac{1}{(\tilde{\mathbf{x}}^\top \tilde{\mathbf{x}})} \right)\right). \tag{3.3}$$

We can calculate this variance except for σ^2. Further, we can consistently estimate σ^2, for example, with the standard formula

$$\hat{\sigma}^2 = \frac{\left\|\tilde{\mathbf{y}} - \tilde{\mathbf{x}}\hat{\beta}^{OLS}\right\|_2^2}{n-2}.$$

Thus, we receive asymptotically valid z-tests for the null-hypothesis H_0 that the model (3.1) holds. We treat the extension to non-Gaussian $\boldsymbol{\epsilon}$ in Section 3.2.2 (for the multivariate case directly). As discussed above, in the presence of confounding, we can have that $\beta^{HOLS} \neq \beta^{OLS}$. In such situations, a test assuming (3.3) will have asymptotic power equal to 1 for correctly rejecting H_0 under some conditions. These asymptotic results are discussed in Section 3.3.1 and Section 3.3.2.

3.2.2 Multivariate regression

We typically want to examine the goodness of fit of a linear model with $p > 1$ covariates. We consider p to be fixed in this section and discuss the case where p is allowed to diverge with n in Section 3.2.3.

We consider the causal model

$$Y \leftarrow \mathbf{X}^\top \boldsymbol{\beta} + \mathcal{E}, \quad \text{where} \quad \mathbf{X} \perp \mathcal{E}, \quad \mathbb{E}[\mathcal{E}] = 0 \quad \text{and} \quad \mathbb{E}\left[\mathcal{E}^2\right] = \sigma^2 < \infty \tag{3.4}$$

with $\mathbf{X} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. Note that $\mathbb{E}[\mathcal{E}] = 0$ can always be enforced by including an intercept in the set of predictors. We assume the according moment matrix $\boldsymbol{\Sigma}^{\mathbf{X}}$ to be invertible. Then, the principal submatrices $\boldsymbol{\Sigma}^{\mathbf{X}}_{-j,-j} \coloneqq \mathbb{E}\left[\mathbf{X}_{-j}\mathbf{X}_{-j}^\top\right]$ are also invertible. We formulate a global null hypothesis that the

model in (3.4) is correct and we denote it by H_0. To make use of the test described for the univariate case, we consider every component $j \in \{1, \dots, p\}$ separately and work with partial regression, see, e.g., Belsley et al. (2005). For the population version, we define

$$\begin{aligned} Z_j &\coloneqq X_j - \mathbf{X}_{-j}^\top \boldsymbol{\gamma}_j, \qquad \text{where} \quad \boldsymbol{\gamma}_j \coloneqq \underset{\mathbf{b}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \mathbb{E}\left[\left(X_j - \mathbf{X}_{-j}^\top \mathbf{b}\right)^2\right] = \left(\boldsymbol{\Sigma}^{\mathbf{X}}_{-j,-j}\right)^{-1} \mathbb{E}[\mathbf{X}_{-j} X_j] \\ W_j &\coloneqq Y - \mathbf{X}_{-j}^\top \boldsymbol{\zeta}_j, \qquad \text{where} \quad \boldsymbol{\zeta}_j \coloneqq \underset{\mathbf{b}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - \mathbf{X}_{-j}^\top \mathbf{b}\right)^2\right] = \left(\boldsymbol{\Sigma}^{\mathbf{X}}_{-j,-j}\right)^{-1} \mathbb{E}[\mathbf{X}_{-j} Y]. \end{aligned} \tag{3.5}$$

Under H_0, it holds that $W_j = Z_j \beta_j + \mathcal{E}$. For $\boldsymbol{\beta}^{OLS} \coloneqq \left(\boldsymbol{\Sigma}^{\mathbf{X}}\right)^{-1} \mathbb{E}[\mathbf{X}Y]$, we find

$$\beta_j^{OLS} = \frac{\mathbb{E}[Z_j W_j]}{\mathbb{E}\left[Z_j^2\right]} \overset{H_0}{=} \beta_j.$$

The first equality is a well-known application of the Frish-Waugh theorem, see, e.g., Greene (2003). We define the according HOLS parameter by partial regression for every component j separately, namely

$$\beta_j^{HOLS} \coloneqq \frac{\mathbb{E}\left[Z_j^3 W_j\right]}{\mathbb{E}\left[Z_j^4\right]} \overset{H_0}{=} \beta_j.$$

We define a local, i.e., per-covariate null hypothesis $H_{0,j}: \ \beta_j^{OLS} = \beta_j^{HOLS}$. The difference $\beta_j^{OLS} - \beta_j^{HOLS}$ can detect certain local alternatives from the null hypothesis H_0. Here, local refers to the covariate X_j whose effect on Y is potentially confounded or involves a nonlinearity. Under model (3.4), $H_{0,j}$ holds true for every j. We discuss in Sections 3.3 and 3.4 some concrete examples, where it is insightful to consider tests for individual $H_{0,j}$.

We turn to sample estimates of the parameters. The residuals are estimated by

$$\begin{aligned} \hat{\mathbf{z}}_j &= \mathbf{x}_j - \boldsymbol{P}_{-j}\mathbf{x}_j = \boldsymbol{P}_{-j}^\perp \mathbf{x}_j \quad \text{and} \\ \hat{\mathbf{w}}_j &= \mathbf{y} - \boldsymbol{P}_{-j}\mathbf{y} = \boldsymbol{P}_{-j}^\perp \mathbf{y} \overset{H_0}{=} \boldsymbol{P}_{-j}^\perp(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \hat{\mathbf{z}}_j \beta_j + \boldsymbol{P}_{-j}^\perp \boldsymbol{\epsilon}. \end{aligned}$$

With ordinary least squares, we receive $\hat{\beta}_j^{OLS}$ from regression of $\hat{\mathbf{w}}_j$ versus $\hat{\mathbf{z}}_j$, where the error term is $\boldsymbol{P}_{-j}^\perp \boldsymbol{\epsilon}$. Accordingly, we calculate $\hat{\beta}_j^{HOLS}$ from regression of $\hat{\mathbf{z}}_j \odot \hat{\mathbf{w}}_j$ versus $\hat{\mathbf{z}}_j^2$ with an error term $\hat{\mathbf{z}}_j \odot \boldsymbol{P}_{-j}^\perp \boldsymbol{\epsilon}$. Thus, we define

$$\hat{\beta}_j^{OLS} \coloneqq \frac{\hat{\mathbf{z}}_j^\top \hat{\mathbf{w}}_j}{\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j} \quad \text{and} \quad \hat{\beta}_j^{HOLS} \coloneqq \frac{\left(\hat{\mathbf{z}}_j^2\right)^\top (\hat{\mathbf{z}}_j \odot \hat{\mathbf{w}}_j)}{\left(\hat{\mathbf{z}}_j^2\right)^\top \left(\hat{\mathbf{z}}_j^2\right)} = \frac{\left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{w}}_j}{\left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{z}}_j}. \tag{3.6}$$

This is analogous to the univariate case, where we have $\tilde{\mathbf{y}}$ instead of $\hat{\mathbf{w}}_j$, $\tilde{\mathbf{x}}$ instead of $\hat{\mathbf{z}}_j$ and $\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^\top\right)$ instead of $\boldsymbol{P}_{-j}^{\perp}$, and $\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^\top\right)$ can be thought of as orthogonal projection onto $\mathbf{e}$'s complement, which completes the analogy. Again, we see that given $\mathbf{x}$, $\left(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right)$ is some known linear combination of $\boldsymbol{\epsilon}$, thus, it is conditionally Gaussian for Gaussian $\boldsymbol{\epsilon}$. The same holds for $\left(\hat{\boldsymbol{\beta}}^{HOLS} - \hat{\boldsymbol{\beta}}^{OLS}\right)$.

Naturally, Gaussian $\mathcal{E}$ is an overly strong assumption. Therefore, we consider additional assumptions such that the central limit theorem can be invoked.

(A3.1) The moment matrix $\boldsymbol{\Sigma}^{\mathbf{X}}$ has positive smallest eigenvalue.

(A3.2) $\mathbb{E}\left[X_j^6\right] < \infty$ and $\mathbb{E}\left[Z_j^6\right] < \infty \; \forall j$.

Further, let

$$\tilde{Z}_j^3 \coloneqq Z_j^3 - \mathbf{X}_{-j}^\top \tilde{\boldsymbol{\gamma}}_j, \text{ where } \tilde{\boldsymbol{\gamma}}_j \coloneqq \underset{\mathbf{b}\in\mathbb{R}^{p-1}}{\operatorname{argmin}} \mathbb{E}\left[\left(Z_j^3 - \mathbf{X}_{-j}^\top \mathbf{b}\right)^2\right] = \left(\boldsymbol{\Sigma}_{-j,-j}^{\mathbf{X}}\right)^{-1} \mathbb{E}[\mathbf{X}_{-j} Z_j^3]. \tag{3.7}$$

Note that $\mathbb{E}\left[\left(\tilde{Z}_j^3\right)^2\right] \leq \mathbb{E}\left[Z_j^6\right] < \infty$.

Theorem 3.1. Assume that the data follows the model (3.4) and that (A3.1) - (A3.2) hold. Let p be fixed and $n \to \infty$. Then,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{HOLS} - \hat{\boldsymbol{\beta}}^{OLS}\right) \xrightarrow{\mathbb{D}} \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbb{E}\left[\mathbf{V}\mathbf{V}^\top\right]\right)$$
$$\frac{1}{n}\hat{\mathbf{v}}^\top \hat{\mathbf{v}} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\mathbf{V}\mathbf{V}^\top\right],$$

where $\hat{\mathbf{v}}_j = \dfrac{\boldsymbol{P}_{-j}^{\perp}\left(\hat{\mathbf{z}}_j^3\right)}{\frac{1}{n}\left(\hat{\mathbf{z}}_j^2\right)^\top \left(\hat{\mathbf{z}}_j^2\right)} - \dfrac{\hat{\mathbf{z}}_j}{\frac{1}{n}\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j}$ and $V_j = \dfrac{\tilde{Z}_j^3}{\mathbb{E}\left[Z_j^4\right]} - \dfrac{Z_j}{\mathbb{E}\left[Z_j^2\right]}$.

Note that

$$\left(\hat{\boldsymbol{\beta}}^{HOLS} - \hat{\boldsymbol{\beta}}^{OLS}\right) \overset{H_0}{=} \frac{1}{n}\hat{\mathbf{v}}^\top \boldsymbol{\epsilon}, \quad \text{and, in analogy to (3.3)}, \quad \frac{1}{n^2}\hat{\mathbf{v}}_j^\top \hat{\mathbf{v}}_j = \frac{\left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^{\perp}\left(\hat{\mathbf{z}}_j^3\right)}{\left(\left(\hat{\mathbf{z}}_j^2\right)^\top \left(\hat{\mathbf{z}}_j^2\right)\right)^2} - \frac{1}{\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j}.$$

Following Theorem 3.1, we can test the null hypothesis H_0 with a consistent estimate for σ^2. Such an estimate can be obtained, e.g., using the standard formula

$$\hat{\sigma}^2 = \frac{\left\|\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2}{n-p}.$$

We define for later reference

$$\widehat{\mathrm{Var}}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big) \coloneqq \hat{\sigma}^2 \frac{1}{n^2} \hat{\mathbf{v}}_j^\top \hat{\mathbf{v}}_j. \tag{3.8}$$

To test $H_{0,j}$, we can compare $\left(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right)$ to the quantiles of the univariate normal distribution with the according variance. The joint distribution leads to a global test that controls the type I error. Namely, one can look at the maximum test statistic $T = \max_k \left|\hat{\beta}_k^{HOLS} - \hat{\beta}_k^{OLS}\right| \overset{H_0}{\sim} \max_k |S_k|$, where $\mathbf{S} \sim \mathcal{N}\big(\mathbf{0}, \hat{\sigma}^2 \hat{\mathbf{v}}^\top \hat{\mathbf{v}}/n^2\big)$ can be easily simulated. Further, one receives multiplicity corrected individual p-values for $H_{0,j}$ by comparing each $\left|\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right|$ to the distribution of $\max_k |S_k|$. This is in analogy to the multiplicity correction suggested by Bühlmann (2013). Naturally, other multiplicity correction techniques such as Bonferroni-Holm are valid as well.

Algorithm 3.1 summarizes how to find both raw and multiplicity corrected p-values for each component j corresponding to the jth covariate, p_j and P_j respectively. Then, one would reject the global null hypothesis H_0 that the model (3.4) holds if $\min_j P_j \leq \alpha$, and such a decision procedure provides control of the type I error at level α. Note that this means that we have strong control of the FWER for testing all $H_{0,j}$.

Algorithm 3.1 HOLS check

1: **for** $j = 1$ to p **do**
2: $\quad \boldsymbol{P}_{-j}^{\perp} = \boldsymbol{I}_n - \mathbf{x}_{-j}\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)\mathbf{x}_{-j}^\top$
3: $\quad$ Regress $\mathbf{x}_j$ versus $\mathbf{x}_{-j}$ via least squares, denote the residual by $\hat{\mathbf{z}}_j = \boldsymbol{P}_{-j}^{\perp}\mathbf{x}_j$
4: $\quad$ Regress $\mathbf{y}$ versus $\mathbf{x}_{-j}$ via least squares, denote the residual by $\hat{\mathbf{w}}_j = \boldsymbol{P}_{-j}^{\perp}\mathbf{y}$
5: $\quad \hat{\beta}_j^{OLS} = \dfrac{\hat{\mathbf{z}}_j^\top \hat{\mathbf{w}}_j}{\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j}$, $\hat{\beta}_j^{HOLS} = \dfrac{\Big(\hat{\mathbf{z}}_j^3\Big)^\top \hat{\mathbf{w}}_j}{\Big(\hat{\mathbf{z}}_j^2\Big)^\top \Big(\hat{\mathbf{z}}_j^2\Big)}$ and $\hat{\mathbf{v}}_j = \dfrac{\boldsymbol{P}_{-j}^{\perp}\Big(\hat{\mathbf{z}}_j^3\Big)}{\dfrac{1}{n}\Big(\hat{\mathbf{z}}_j^2\Big)^\top \Big(\hat{\mathbf{z}}_j^2\Big)} - \dfrac{\hat{\mathbf{z}}_j}{\dfrac{1}{n}\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j}$
6: $\hat{\sigma}^2 = \dfrac{\big\|\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}^{OLS}\big\|_2^2}{n-p}$
7: Create n_{sim} i.i.d copies of $\mathbf{S} \sim \mathcal{N}\big(\mathbf{0}, \hat{\sigma}^2 \hat{\mathbf{v}}^\top \hat{\mathbf{v}}/n^2\big)$, say, $\mathbf{s}^1$ to $\mathbf{s}^{n_{sim}}$
8: **for** $j = 1$ to p **do**
9: $\quad p_j = 2\left(1 - \Phi\left(\dfrac{\left|\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right|}{\hat{\sigma}\dfrac{1}{n}\|\hat{\mathbf{v}}_j\|_2}\right)\right)$
10: $\quad P_j = \dfrac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{1}\Big(\left|\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right| > \big\|\mathbf{s}^i\big\|_\infty\Big)$

Corollary 3.1. Assume the conditions of Theorem 3.1. Consider the decision rule to reject H_0 iff $\min_j P_j \leq \alpha$, where P_j is as in Step 10 of Algorithm 3.1. Then, the type I error is asymptotically controlled at α. Furthermore, the FWER is asymptotically controlled at level α for testing all local hypotheses $\{H_{0,j};\ j = 1, \ldots, p\}$ with the decision rule to reject $H_{0,j}$ iff $P_j \leq \alpha$.

We provide simulation results supporting this theory in Section 3.A of the supplemental material.

3.2.3 High-dimensional data

We now turn to high-dimensional situations. We assume the global null hypothesis (3.4) but allow for p to diverge with and even exceed n such that ordinary least squares regression is not applicable. Instead, we apply the debiased Lasso introduced in Zhang and Zhang (2014) and further discussed in van de Geer et al. (2014). We denote the estimator again by $\hat{\boldsymbol{\beta}}^{OLS}$ since it converges under certain conditions to the population parameter $\boldsymbol{\beta}^{OLS}$.

From the debiased Lasso, we receive $\hat{\mathbf{z}}_j = \mathbf{x}_j - \mathbf{x}_{-j}\hat{\boldsymbol{\gamma}}_j$, where $\hat{\boldsymbol{\gamma}}_j$ is obtained by regressing $\mathbf{x}_j$ versus $\mathbf{x}_{-j}$ using the Lasso, and $\hat{\mathbf{w}}_j = \mathbf{y} - \mathbf{x}_{-j}\hat{\boldsymbol{\beta}}_{-j}$ with $\hat{\boldsymbol{\beta}}$ coming from the Lasso fit of $\mathbf{y}$ versus $\mathbf{x}$. Since $\hat{\beta}_j^{OLS} = \hat{\mathbf{z}}_j^\top \hat{\mathbf{w}}_j / \hat{\mathbf{z}}_j^\top \mathbf{x}_j$, one might want to use $\left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{w}}_j / \left(\hat{\mathbf{z}}_j^3\right)^\top \mathbf{x}_j$ for HOLS. However, this leads in general to an uncontrollable approximation error since $\mathbb{E}\left[Z_j^3 \mathbf{X}_{-j}\right] \neq \mathbf{0}$. As a remedy, we suggest a second level of orthogonalization based on $\tilde{Z}_j^3$ and $\tilde{\boldsymbol{\gamma}}_j$ as defined in (3.7). Naturally, we have $\tilde{Z}_j^3 = Z_j^3$ iff $\mathbb{E}\left[Z_j^3 \mathbf{X}_{-j}\right] = \mathbf{0}$ and always $\mathbb{E}\left[\tilde{Z}_j^3 \mathbf{X}_{-j}\right] = \mathbf{0}$. To approximate $\tilde{\mathbf{z}}_j^3$ we use the Lasso for the regression $\hat{\mathbf{z}}_j^3$ versus $\mathbf{x}_{-j}$ leading to $\hat{\tilde{\mathbf{z}}}_j^3 = \hat{\mathbf{z}}_j^3 - \mathbf{x}_{-j}\hat{\tilde{\boldsymbol{\gamma}}}_j$. We define $\hat{\boldsymbol{\beta}}^{HOLS}$ as

$$\hat{\beta}_j^{HOLS} := \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \hat{\mathbf{w}}_j}{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_j} = \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \left(\mathbf{y} - \mathbf{x}_{-j}\hat{\boldsymbol{\beta}}_{-j}\right)}{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_j} \overset{H_0}{=} \beta_j + \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_{-j}\left(\boldsymbol{\beta}_{-j} - \hat{\boldsymbol{\beta}}_{-j}\right)/n}{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_j/n} + \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \boldsymbol{\epsilon}}{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_j}.$$

Finally, we are interested in the difference between $\hat{\beta}_j^{HOLS}$ and $\hat{\beta}_j^{OLS}$. Under suitable assumptions for the sparsity, the moment matrix, and the tail behaviour of $\mathbf{X}$ and $\mathcal{E}$, we can derive the limiting Gaussian distribution of this difference allowing for asymptotically valid tests. We apply Algorithm 3.2 where we make use of the (asymptotic) normality of the non-vanishing term in this difference. For non-Gaussian $\mathcal{E}$, a multiplicity correction method that does not rely on exact Gaussianity of this remainder might be preferred since the CLT does not apply for dimensions growing too fast.

We provide here the main result to justify Algorithm 3.2 invoking additional assumptions on the dimensionality and sparsity of the problem. We use the definitions $s := \|\boldsymbol{\beta}\|_0$, $s_j := \left\|\boldsymbol{\gamma}_j\right\|_0$ and $\tilde{s}_j := \left\|\tilde{\boldsymbol{\gamma}}_j\right\|_0$ to denote the different levels of sparsity.

(C3.1) The design matrix $\mathbf{x}$ has i.i.d. sub-Gaussian rows. The moment matrix $\boldsymbol{\Sigma}^{\mathbf{X}}$ has strictly positive smallest eigenvalue $\Lambda_{\min}^2$ satisfying $1/\Lambda_{\min}^2 = \mathcal{O}(1)$. Also, $\max_j \boldsymbol{\Sigma}_{j,j}^{\mathbf{X}} = \mathcal{O}(1)$.

(C3.2) $s = \mathcal{O}\left(\frac{n^{1/2}}{\log(p)^3}\right)$.

(C3.3) $ss_j^2 = \mathcal{O}\left(\frac{n^{3/2}}{\log(p)^3}\right)$, $ss_j = \mathcal{O}\left(\frac{n}{\log(p)^{5/2}}\right)$ and $ss_j^{1/2} = \mathcal{O}\left(\frac{n^{1/2}}{\log(p)^{3/2}}\right)$.

(C3.4) $s_j = \mathcal{O}\left(\frac{n^{3/5}}{\log(p)}\right)$. **(C3.5)** $\sqrt{n}s\lambda\tilde{\lambda}_j = \mathcal{O}(1)$. **(C3.6)** $\tilde{s}_j\tilde{\lambda}_j^2 = \mathcal{O}(1)$.

Theorem 3.2. Assume that the data follows the model (3.4) with sub-Gaussian $\mathcal{E}$ and that (C3.1) - (C3.6) hold ($\forall j$). Let $\hat{\boldsymbol{\beta}}$ come from Lasso regression with $\lambda \asymp \sqrt{\log(p)/n}$, $\hat{\mathbf{z}}_j$ from nodewise Lasso regression using $\lambda_j \asymp \sqrt{\log(p)/n}$, and $\hat{\tilde{\mathbf{z}}}_j^3$ from nodewise Lasso regression of $\hat{\mathbf{z}}_j^3$ versus $\mathbf{x}_{-j}$ using $\tilde{\lambda}_j \asymp \max\left\{\log(p)^{5/2}n^{-1/2}, s_j^2\log(p)^{5/2}n^{-3/2}, s_j\log(p)^2n^{-1}, \sqrt{s_j}\log(p)n^{-1/2}\right\}$. Let $\hat{\sigma}$ be any consistent estimator for σ. Then,

$$\frac{\sqrt{n}\left(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right)}{\sqrt{\hat{\sigma}^2\frac{1}{n}\hat{\mathbf{v}}_j^\top\hat{\mathbf{v}}_j}} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1) \quad \text{where} \quad \hat{\mathbf{v}}_j = \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)}{\frac{1}{n}\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top\mathbf{x}_j} - \frac{\hat{\mathbf{z}}_j}{\frac{1}{n}\hat{\mathbf{z}}_j^\top\mathbf{x}_j}.$$

Algorithm 3.2 HOLS check for $p > n$

1: Regress $\mathbf{y}$ versus $\mathbf{x}$ via Lasso with a penalty parameter λ, denote the estimated regression coefficients by $\hat{\boldsymbol{\beta}}$
2: **for** $j = 1$ to p **do**
3: Regress $\mathbf{x}_j$ versus $\mathbf{x}_{-j}$ via Lasso with a penalty parameter λ_j, denote the residual by $\hat{\mathbf{z}}_j$
4: Regress $\hat{\mathbf{z}}_j^3$ versus $\mathbf{x}_{-j}$ via Lasso with a penalty parameter $\tilde{\lambda}_j$, denote the residual by $\hat{\tilde{\mathbf{z}}}_j^3$
5: $\hat{\mathbf{w}}_j = \mathbf{y} - \mathbf{x}_{-j}\hat{\boldsymbol{\beta}}_{-j}$
6: $\hat{\beta}_j^{OLS} = \frac{\hat{\mathbf{z}}_j^\top\hat{\mathbf{w}}_j}{\hat{\mathbf{z}}_j^\top\mathbf{x}_j}$, $\hat{\beta}_j^{HOLS} = \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top\hat{\mathbf{w}}_j}{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top\mathbf{x}_j}$ and $\hat{\mathbf{v}}_j = \frac{\left(\hat{\tilde{\mathbf{z}}}_j^3\right)}{\frac{1}{n}\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top\mathbf{x}_j} - \frac{\hat{\mathbf{z}}_j}{\frac{1}{n}\hat{\mathbf{z}}_j^\top\mathbf{x}_j}$
7: $\hat{\sigma}^2 = \frac{\left\|\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}\right\|_2^2}{n - \left|\hat{\boldsymbol{\beta}}\right|_0}$ (or any other reasonable variance estimator)
8: Create n_{sim} i.i.d copies of $\mathbf{S} \sim \mathcal{N}\left(\mathbf{0}, \hat{\sigma}^2\frac{1}{n^2}\hat{\mathbf{v}}^\top\hat{\mathbf{v}}\right)$, say, $\mathbf{s}^1$ to $\mathbf{s}^{n_{sim}}$
9: **for** $j = 1$ to p **do**
10: $p_j = 2\left(1 - \Phi\left(\frac{\left|\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right|}{\hat{\sigma}\frac{1}{n}\|\hat{\mathbf{v}}_j\|_2}\right)\right)$
11: $P_j = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\mathbb{1}\left(\left|\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right| > \left\|\mathbf{s}^i\right\|_\infty\right)$

We defer the technical details to Section 3.C of the supplemental material. Simulation results concerning high-dimensional data can be found in Section 3.A of the supplemental material.

3.3 The confounded case and local null hypotheses

In this section and the following, we mainly exploit confounding in linear models as the alternative hypothesis since these are the models where our tests for the local null hypotheses $H_{0,j}$ are most informative. For a discussion of which interpretations might carry over to more general data generating distributions, we refer to Section 3.4.3.

Note that everything that is discussed in Sections 3.3 and 3.4 implicitly applies to high-dimensional data as well under suitable assumptions. We refrain from going into detail for the sake of brevity. Thus, Theorems 3.3 - 3.6 which contain our main asymptotic results for the local intepretation are designed explicitly for the fixed p case.

We look at the causal model

$$\begin{aligned} \mathbf{X} &\leftarrow \boldsymbol{\rho}\mathbf{H} + \mathcal{E}_{\mathbf{X}} \\ Y &\leftarrow \mathbf{X}^\top \boldsymbol{\beta} + \mathbf{H}^\top \boldsymbol{\alpha} + \mathcal{E}, \end{aligned} \tag{3.9}$$

where $\mathbf{H} \in \mathbb{R}^d$, $\mathcal{E}_{\mathbf{X}} \in \mathbb{R}^p$ and $\mathcal{E} \in \mathbb{R}$ are independent and centered random variables, and $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $\boldsymbol{\rho} \in \mathbb{R}^{p \times d}$ are fixed model parameters. Thus, there exists some hidden confounder $\mathbf{H}$. For the inner product matrices, it holds that

$$\boldsymbol{\Sigma}^{\mathbf{X}} = \boldsymbol{\Sigma}^{\mathcal{E}_{\mathbf{X}}} + \boldsymbol{\rho}\boldsymbol{\Sigma}^{\mathbf{H}}\boldsymbol{\rho}^\top.$$

Furthermore, we have

$$\boldsymbol{\beta}^{OLS} = \left(\boldsymbol{\Sigma}^{\mathbf{X}}\right)^{-1}\mathbf{E}[\mathbf{X}Y] = \left(\boldsymbol{\Sigma}^{\mathbf{X}}\right)^{-1}\left(\boldsymbol{\Sigma}^{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\rho}\boldsymbol{\Sigma}^{\mathbf{H}}\boldsymbol{\alpha}\right) = \boldsymbol{\beta} + \left(\boldsymbol{\Sigma}^{\mathbf{X}}\right)^{-1}\boldsymbol{\rho}\boldsymbol{\Sigma}^{\mathbf{H}}\boldsymbol{\alpha} \tag{3.10}$$

We will generally refer to $\beta_j^{OLS} \neq \beta_j$, where β_j is according to model (3.9), as confounding bias on β_j^{OLS}. Further, when writing directly confounded, we mean covariate indices j for which $X_j \neq \mathcal{E}_{X_j}$.

Note that we can always decompose Y both globally and locally as follows

$$Y = \mathbf{X}^\top \boldsymbol{\beta}^{OLS} + \tilde{\mathcal{E}}, \quad \text{with} \quad \mathbb{E}\left[\mathbf{X}\tilde{\mathcal{E}}\right] = \mathbf{0}, \quad \mathbb{E}\left[\tilde{\mathcal{E}}\right] = 0 \quad \text{but (potentially)} \quad \mathbf{X} \not\perp \tilde{\mathcal{E}} \tag{3.11}$$

$$W_j = Z_j \beta_j^{OLS} + \tilde{\mathcal{E}}, \quad \text{with} \quad \mathbb{E}\left[Z_j\tilde{\mathcal{E}}\right] = \mathbf{0}, \quad \mathbb{E}\left[\tilde{\mathcal{E}}\right] = 0 \quad \text{but (potentially)} \quad Z_j \not\perp \tilde{\mathcal{E}} \tag{3.12}$$

using the definitions from (3.5). We now want to see how $\boldsymbol{\beta}^{OLS}$ relates to $\boldsymbol{\beta}$ in certain models. Especially, we are interested in whether there is some potential local interpretation in the sense of distinguishing between "confounded" and "unconfounded" variables. From (3.10), we see that this is

linked to the structure of the covariance matrices as well as $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$. We define the sets

$$V = \{j : \beta_j^{OLS} = \beta_j\} \quad \text{and} \quad U = \{j : \beta_j^{OLS} = \beta_j^{HOLS}\} = \{j : H_{0,j} \text{ is true}\}. \tag{3.13}$$

Using the Woodbury matrix identity, we find a sufficient condition

$$j \in V \quad \text{if} \quad \boldsymbol{\rho}^\top \left(\boldsymbol{\Sigma}^{\mathcal{E}\mathbf{x}}\right)_j^{-1} = \mathbf{0} \quad \text{which is implied by}$$
$$\left\{k \in \{1,\dots,p\} : \ \left(\boldsymbol{\Sigma}^{\mathcal{E}\mathbf{x}}\right)_{jk}^{-1} \neq 0\right\} \cap \left\{l \in \{1,\dots,p\} : \ \left\|\mathbf{e}_l^\top \boldsymbol{\rho}\right\| > 0\right\} = \emptyset. \tag{3.14}$$

Thus, if the intersection between covariates that have linear predictive power for X_j and covariates that are directly confounded is empty, it must hold that $\beta_j^{OLS} = \beta_j$. Therefore, we can indeed say that for these variables we estimate the true causal effect using ordinary least squares.

To correctly detect V, we would like $\beta_j^{HOLS} = \beta_j^{OLS} = \beta_j$. As β_j^{HOLS} involves higher-order moments, knowledge of the covariance structure is not sufficient to check this. From (3.12), we see that $\mathbb{E}\left[Z_j^3 \tilde{\mathcal{E}}\right] = 0$ is necessary and sufficient to ensure $j \in U$. In Section 3.3.2, we discuss the two cases where detection fails, i.e., $U \setminus V \neq \emptyset$ and $V \setminus U \neq \emptyset$. We present models for which we can characterize a set of variables which are in $U \cap V = \left\{j : \beta_j^{HOLS} = \beta_j^{OLS} = \beta_j\right\}$ in Section 3.4.

3.3.1 Sample estimates

For a confounded model, the hope is that the global test $\min_j P_j \leq \alpha$, where P_j is the adjusted p-value according to Step 10 in Algorithm 3.1, leads to a rejection of H_0, i.e., the modelling assumption (3.4). One could further examine the local structure and, based on the corrected p-values P_j, distinguish the predictors for which we have evidence that $\beta_j^{HOLS} \neq \beta_j^{OLS}$. We consider in the following this local interpretation, showing that we asymptotically control the type-I error and receive power approaching 1. Implicitly, we assume that U is a useful proxy for V.

For all asymptotic results in this section, we assume p to be fixed and $n \to \infty$ as in Theorem 3.1.

Theorem 3.3. Assume that the data follows the model (3.11) and that (A3.1)-(A3.2) hold. Assume further $\sigma_{\tilde{\mathcal{E}}}^2 = \mathbb{E}\left[\tilde{\mathcal{E}}^2\right] < \infty$. Then,

$$\hat{\beta}_j^{OLS} = \beta_j^{OLS} + \mathcal{O}_p(1), \quad \hat{\beta}_j^{HOLS} = \beta_j^{HOLS} + \mathcal{O}_p(1) \quad \text{and} \quad \widehat{\text{Var}}\left(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right) = \mathcal{O}_p\left(\frac{1}{n}\right),$$

where $\widehat{\text{Var}}\left(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right)$ is according to (3.8).

Thus, for some fixed alternative $\left|\beta_j^{HOLS} - \beta_j^{OLS}\right| > 0$, the absolute z-statistics increases as $\sqrt{n}$.

In order to get some local interpretation, the behaviour for variables $j \in U$ is of importance. If $\left|\beta_j^{HOLS} - \beta_j^{OLS}\right| = 0$, Theorem 3.3 is not sufficient to understand the asymptotic behaviour. We

refine the results using additional assumptions.

(A3.3) $\mathbb{E}\Big[\Big(X_j\tilde{\mathcal{E}}\Big)^2\Big] < \infty \; \forall j$ **(B3.2)** $Z_j \perp \tilde{\mathcal{E}}$

(B3.1) $\mathbb{E}\Big[Z_j^2 X_k \tilde{\mathcal{E}}\Big] = 0 \; \forall k \neq j$ **(B3.3)** $\tilde{Z}_j^3 \perp \tilde{\mathcal{E}}$

Note that we use different letters for the assumptions to distinguish between those that are essentially some (mild) moment conditions and those that truly make nodes unconfounded. Obviously, (B3.2) is not necessary for $\beta_j^{OLS} = \beta_j^{HOLS}$, but we will focus on these variables as these are the ones that are truly unconfounded in the sense that the projected single variable model (3.12) has an independent error term, while as for other variables it can be rather considered an unwanted artefact of our method. Furthermore, the derived asymptotic variance results only hold true when assuming (B3.2) and (B3.3) as well. Assumption (A3.3) implies a further moment condition. Especially, when considering nonlinearities, there exist cases for which (A3.3) is not implied by (A3.2). We discuss Assumptions (B3.1), (B3.2) and (B3.3) for certain models in Section 3.4. Condition (B3.1) seems to be a bit artificial but is invoked in the proofs. We argue in Section 3.4 that it is naturally linked to the models in scope.

Theorem 3.4. Assume that the data follows the model (3.11) and that (A3.1) - (A3.3) hold. Let j be some covariate with $\beta_j^{OLS} = \beta_j^{HOLS}$ for which (B3.1) - (B3.3) hold. Then,

$$\frac{\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big)}{\sqrt{\widehat{\mathrm{Var}}\Big(\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big)\Big)}} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1).$$

Thus, for these predictors we receive asymptotically valid tests.

Multiplicity correction In order not to falsely reject the local null hypothesis $H_{0,j}$ for any covariate with $j \in U$ (with probability at least $1-\alpha$), we need to invoke some multiplicity correction. Analogously to Section 3.2.2, one can see that $\hat{\boldsymbol{\beta}}^{HOLS} - \hat{\boldsymbol{\beta}}^{OLS} = \hat{\mathbf{v}}^\top \tilde{\boldsymbol{\epsilon}}/n$, which enables the multiplicity correction as in Algorithm 3.1.

Theorem 3.5. Assume that the data follows the model (3.11) and that (A3.1) - (A3.3) hold. Let U' be the set of variables j for which $j \in U$ and (B3.1) - (B3.3) hold. Then,

$$\sqrt{n}\Big(\hat{\boldsymbol{\beta}}_{U'}^{HOLS} - \hat{\boldsymbol{\beta}}_{U'}^{OLS}\Big) \xrightarrow{\mathbb{D}} \mathcal{N}\Big(\mathbf{0}, \sigma_{\tilde{\mathcal{E}}}^2 \mathbb{E}\Big[\mathbf{V}_{U'}\mathbf{V}_{U'}^\top\Big]\Big)$$
$$\frac{1}{n}\hat{\mathbf{v}}_{U'}^\top \hat{\mathbf{v}}_{U'} \xrightarrow{\mathbb{P}} \mathbb{E}\Big[\mathbf{V}_{U'}\mathbf{V}_{U'}^\top\Big]$$

Corollary 3.2. Assume the conditions of Theorem 3.5. Consider the decision rule to reject $H_{0,j}$ iff $P_j \leq \alpha$, where P_j is as in Step 10 of Algorithm 3.1. Then, the familywise error rate amongst the set U' is asymptotically controlled at α.

3.3.2 Inferring V from U

Recall the definitions in (3.13). U is the set that we try to infer with our HOLS check. Naturally, one would rather be interested in the set V, which consists of the variables for which we can consistently estimate the true linear causal effect according to (3.9) through linear regression. We discuss here when using U as proxy for V might fail and especially analyse how variables could belong to the difference between the sets. For this, recall our formulation of the model when the global null hypothesis does not hold true in (3.11) and (3.12). Note that $j \in U$ is equivalent to $\mathbb{E}\Big[Z_j^3 \tilde{\mathcal{E}}\Big] = 0$.

For any variable $j \in U \setminus V$, certain modelling assumptions, that we discuss in the sequel, cannot be fulfilled but they are not necessary for $\mathbb{E}\Big[Z_j^3 \tilde{\mathcal{E}}\Big] = 0$. Especially, the last equality always holds if both $\mathcal{E}_{\mathbf{X}}$ and H jointly have Gaussian kurtosis. If they are even jointly Gaussian, then it is clear that $\mathbf{X} \perp \tilde{\mathcal{E}}$ such that the model (3.11) has independent Gaussian error. Thus, when using only observational data, it behaves exactly like a model under the global null hypothesis and, naturally, we cannot infer the confounding effect. Apart from Gaussian kurtosis, $j \in U \setminus V$ would be mainly due to special constellations implying cancellation of terms that one does not expect to encounter in normal circumstances.

For $j \in V \setminus U$, Z_j and $\tilde{\mathcal{E}}$ must not be independent. As $Z_j \not\perp \tilde{\mathcal{E}}$, the single-covariate model (3.12) is not a linear causal model with independent error term as given in (3.1). Therefore, from a causal inference perspective, one can argue that rejecting the local null hypothesis $H_{0,j}$ is the right thing to do in this case. Furthermore, having variables $j \in V$ is usually related to certain model assumptions except for very specific data setups that lead to cancellation of terms. Under these assumptions, $Z_j \perp \tilde{\mathcal{E}}$ is then usually implied. An example where $\boldsymbol{\beta}^{OLS} = \boldsymbol{\beta}$, but (potentially) $\mathbf{X} \not\perp \tilde{\mathcal{E}}$ is data for which $\boldsymbol{\rho}\boldsymbol{\Sigma}^{\mathbf{H}}\boldsymbol{\alpha} = 0$ using the definitions from model (3.9).

Recovery of U Based on our asymptotic results when the global null does not hold true, we would like to construct a method that perfectly detects the unconfounded variables as $n \to \infty$. Define

$$\hat{U} = \{j : H_{0,j} \text{ not rejected}\} \tag{3.15}$$

The question is how and when can we ensure that

$$\lim_{n\to\infty} \mathbb{P}\Big[\hat{U} = U\Big] = 1.$$

Suppose that we conduct our local z-tests at level α_n, which varies with the sample size such that $\alpha_n \to 0$ as $n \to \infty$. It will be more convenient to interpret this as a threshold on the (scaled) absolute z-statistics, say, τ_n that grows with n, where the z-statistics is defined as

$$t_j = \frac{\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big)}{\sqrt{\widehat{\mathrm{Var}}\Big(\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big)\Big)}}.$$

We refrain from calling it z_j to avoid confusion. We use an additional assumption which is a relaxed version of (B3.3).

(A3.4) $\mathbb{E}\Big[\Big(\tilde{Z}_j^3\tilde{\mathcal{E}}\Big)^2\Big] < \infty$

Theorem 3.6. Assume that the data follows the model (3.11) and that (A3.1) - (A3.3) hold. Assume that (B3.1) and (A3.4) hold $\forall j \in U$. Let τ_n be the threshold on the absolute z-statistics to reject the according null hypothesis with $\tau_n = \mathcal{O}(\sqrt{n})$ and $1/\tau_n = \mathcal{O}(1)$. Then,

$$\lim_{n\to\infty} \mathbb{P}\Big[\hat{U} = U\Big] = 1.$$

In other words, we can choose τ_n to grow at any rate slower than $\sqrt{n}$.

3.4 Specific models

In this section, we discuss two types of models where the local interpretation applies. In these settings, there are variables for which $\beta_j = \beta_j^{OLS} = \beta_j^{HOLS}$ and assumptions (B3.1)-(B3.3) hold even though the overall data follows the model (3.9). We note here first that the model of jointly Gaussian $\mathcal{E}_{\mathbf{X}}$, for which the method is suited, is a special case of the model in Section 3.4.2 below.

3.4.1 Block independence of $\mathcal{E}_{\mathbf{X}}$

Assume that the errors $\mathcal{E}_{\mathbf{X}}$ can be grouped into two or more independent and disjoint blocks. Denote the block that includes j by $B(j)$. Then, it is clear that $\big(\mathbf{\Sigma}^{\mathcal{E}_{\mathbf{X}}}\big)^{-1}_{jk} = 0$ if $B(j) \neq B(k)$. If $\mathbf{X}_{B(j)} = \mathcal{E}_{\mathbf{X}_{B(j)}}$, i.e., the confounder has no effect onto $\mathbf{X}_{B(j)}$, (3.14) holds for all covariates in $B(j)$. Then, no variable in $\mathbf{X}_{B(j)}$ contributes to the best linear predictor for $\mathbf{H}^\top\boldsymbol{\alpha}$. Due to the block independence, this yields $\mathbf{X}_{B(j)} \perp \tilde{\mathcal{E}}$ and $Z_j \perp \tilde{\mathcal{E}}$, i.e., (B3.2) is fulfilled. This also ensures $\mathbb{E}\Big[Z_j^3\tilde{\mathcal{E}}\Big] = 0$. We consider the remaining assumptions: Naturally, the regression Z_j^3 versus $\mathbf{X}_{-j}$ only involves $\mathbf{X}_{B(j)\setminus j}$ and (B3.3) holds as well. For (B3.1), separately consider the case $k \in B(j)$ and $k \notin B(j)$. In the first case, $\mathbb{E}\Big[Z_j^2X_k\tilde{\mathcal{E}}\Big] = \mathbb{E}\Big[Z_j^2X_k\Big]\mathbb{E}\Big[\tilde{\mathcal{E}}\Big] = 0$. In the second case, $\mathbb{E}\Big[Z_j^2X_k\tilde{\mathcal{E}}\Big] = \mathbb{E}\Big[Z_j^2\Big]\mathbb{E}\Big[X_k\tilde{\mathcal{E}}\Big] = 0$.

Theorem 3.7. Assume the data follows the model (3.9) with errors $\mathcal{E}_{\mathbf{X}}$ that can be grouped into independent blocks. Then,

$$\begin{aligned} \beta_j^{HOLS} = \beta_j^{OLS} = \beta_j \ \forall j \quad &\text{for which} \quad \mathbf{X}_{B(j)} = \mathcal{E}_{\mathbf{X}_{B(j)}}. \text{ Further,} \\ \text{(B3.1) -(B3.3) hold } \forall j \quad &\text{for which} \quad \mathbf{X}_{B(j)} = \mathcal{E}_{\mathbf{X}_{B(j)}}. \end{aligned}$$

In some cases, block independence may be a restrictive assumption. Testing this assumption is not an easy problem, and will remain out of the scope of this paper. However, the HOLS check still provides an indirect check of such an assumption since HOLS would likely reject the local null-hypotheses for all covariates, at least for large data-sets, if there is no block that is unaffected by the confounding.

3.4.2 Linear structural equation model

From the previous sections, we know that locally unconfounded structures, in the sense that $\beta_j^{OLS} = \beta_j$, are strongly related to zeroes in the precision matrix. Thus, the question arises for what type of models having zeroes in the precision matrix is a usual thing. Besides block independence, which we have discussed in Section 3.4.1, this will mainly be the case if the data follows a linear structural equation model (SEM). Thus, we will focus on these linear SEMs for the interpretation of local, i.e., by parameter, null hypotheses.

To start, assume that there are no hidden variables. So, let $\mathbf{X}$ be given by the following linear SEM

$$X_j \leftarrow \Psi_j + \sum_{k \in \mathrm{PA}(j)} \theta_{j,k} X_k \quad j = 1, \ldots, p, \tag{3.16}$$

where the $\Psi_1, \ldots, \Psi_p$ are independent and centered random variables. We use the notation $\mathrm{PA}(j)$, $\mathrm{CH}(j)$ and $\mathrm{AN}(j)$ for j's parents, children and ancestors. Further, assume that there exists a directed acyclic graph (DAG) representing this structure. For this type of model, we know that a variable's Markov boundary consists of its parents, its children, and its children's other parents. For every other variable k outside of j's Markov boundary, we have $\left(\boldsymbol{\Sigma}^{\mathbf{X}}\right)^{-1}_{jk} = 0$. Thus, these 0 partial correlations are very usual. In the following, we will analyse how our local tests are especially applicable to this structure.

In the context of linear SEMs, hidden linear confounders can be thought of as unmeasured variables. Therefore, we split $\mathbf{X}$ which contains all possible predictors into two parts. Let $\mathbf{X}_M$ be the measured variables and $\mathbf{X}_N$ the hidden confounder variables. Let $\boldsymbol{\Psi} = \left(\Psi_1, \ldots, \Psi_p\right)^\top$ with the according subsets $\boldsymbol{\Psi}_M$ and $\boldsymbol{\Psi}_N$. Then, we can write

$$\mathbf{X} = \boldsymbol{\omega}\boldsymbol{\Psi}, \quad \mathbf{X}_M = \boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M + \boldsymbol{\omega}_{M,N}\boldsymbol{\Psi}_N \quad \text{and} \quad \mathbf{X}_N = \boldsymbol{\omega}_{N,M}\boldsymbol{\Psi}_M + \boldsymbol{\omega}_{N,N}\boldsymbol{\Psi}_N$$

for some suitable $\boldsymbol{\omega} \in \mathbb{R}^{p\times p}$, where $\omega_{k,l} = 0$ for $k \neq l$ if $l \notin \mathrm{AN}(k)$ and $\omega_{k,k} = 1$. Note that $\boldsymbol{\omega}_{M,M}$ is always invertible since it can be written as a triangular matrix with ones on the diagonal if permuted properly. Under model (3.9), Y can be thought of as a sink node in (3.16). To avoid confusion, we call the parameter if all predictors were observed $\boldsymbol{\beta}^*$. This leads to the definitions

$$
\begin{aligned}
\mathcal{E}_{\mathbf{X}} &\coloneqq \boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M, \quad \boldsymbol{\rho} \coloneqq \boldsymbol{\omega}_{M,N} \quad \text{and} \quad \mathbf{H} \coloneqq \boldsymbol{\Psi}_N \quad \text{such that} \\
\mathbf{X}_M &= \mathcal{E}_{\mathbf{X}} + \boldsymbol{\rho}\mathbf{H} \quad \text{with} \quad \mathcal{E}_{\mathbf{X}} \perp \mathbf{H} \\
Y - \mathcal{E} &= \mathbf{X}^\top\boldsymbol{\beta}^* = \mathbf{X}_M^\top\boldsymbol{\beta}_M^* + \mathbf{X}_N^\top\boldsymbol{\beta}_N^* \\
&= \mathbf{X}_M^\top\left(\boldsymbol{\beta}_M^* + \left(\boldsymbol{\omega}_{N,M}\boldsymbol{\omega}_{M,M}^{-1}\right)^\top\boldsymbol{\beta}_N^*\right) + \mathbf{H}^\top\left(\boldsymbol{\omega}_{N,N} - \boldsymbol{\omega}_{N,M}\boldsymbol{\omega}_{M,M}^{-1}\boldsymbol{\omega}_{M,N}\right)^\top\boldsymbol{\beta}_N^* \\
&\coloneqq \mathbf{X}_M^\top\boldsymbol{\beta} + \mathbf{H}^\top\boldsymbol{\alpha}.
\end{aligned}
\tag{3.17}
$$

When only the given subset is observed we are interested in the parameter $\boldsymbol{\beta}$ as before. We have $\beta_j = \beta_j^*$ iff $\left(\left(\boldsymbol{\omega}_{N,M}\boldsymbol{\omega}_{M,M}^{-1}\right)^\top\boldsymbol{\beta}_N^*\right)_j = 0$.

Theorem 3.8. Assume that the data follows the model (3.16) and (3.17). Let $\mathbf{X}_M$ and $\mathbf{X}_N$ be the observed and hidden variables. Denote by $\mathrm{PA}^M(k)$ the closest ancestors of k that are in M. Consider some $j \in M$.

$$
\text{If} \quad \nexists k \in N : \big(\ j \in \mathrm{PA}^M(k) \text{ and } \beta_k \neq 0\big), \quad \text{then} \quad \beta_j = \beta_j^*.
$$

In other words, the causal parameter can only change for variables that have at least one direct descendant in the hidden set which is a parent of Y itself. By direct descendant, we mean that there is a path from j to k that does not pass any other observed variable. We analyse for which variables we can reconstruct this causal parameter using ordinary least squares regression.

Theorem 3.9. Assume that the data follows the model (3.16) and (3.17). Let $\mathbf{X}_M$ and $\mathbf{X}_N$ be the observed and hidden variables. Then,

$$
\begin{aligned}
\beta_j^{HOLS} = \beta_j^{OLS} = \beta_j \ \forall j \in M \quad &\text{that are not in the Markov boundary of any hidden variable.} \\
\text{(B3.1) -(B3.3) hold } \forall j \in M \quad &\text{that are not in the Markov boundary of any hidden variable.}
\end{aligned}
$$

Thus, for those variables, we can a) correctly retrieve the causal parameter using ordinary least squares regression and b) detect that this is the true parameter by comparing it to β_j^{HOLS}.

Simulation example We assess the performance of our HOLS method in a linear SEM using a simple example. In Figure 3.1, we show the DAG that represents the setup.

For simplicity, the parameters are set such that X_1 to X_7 all have unit variance. X_3 is the only parent of Y and we apply the HOLS method using all but X_3 as predictors, i.e., X_3 is treated as

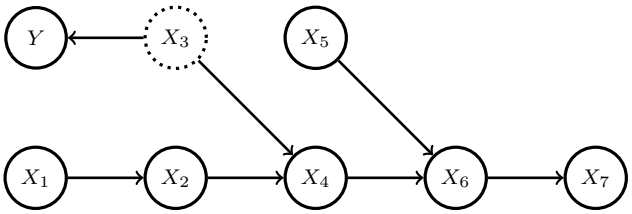


Figure 3.1: DAG of the linear SEM. X_3 is assumed to be hidden which is depicted by the dashed circle. We use the following specifications: $\Psi_1 \overset{\mathbb{D}}{=} \Psi_3 \overset{\mathbb{D}}{=} \Psi_5 \sim t_7/\sqrt{7/5}$, $\Psi_2 \overset{\mathbb{D}}{=} \Psi_6 \overset{\mathbb{D}}{=} \Psi_7 \sim \mathcal{N}(0, 1/2)$, $\Psi_4 \sim \text{Unif}\left[-\sqrt{3/2}, \sqrt{3/2}\right]$ and $\mathcal{E} \sim \mathcal{N}(0, 1)$. $\theta_{2,1} = \theta_{7,6} = \sqrt{1/2}$, $\theta_{4,2} = \theta_{4,3} = \theta_{6,4} = \theta_{6,5} = 0.5$ and $\beta_3^* = \sqrt{5/2}$.

hidden variable. Following Theorem 3.9, we know that for variables X_1 and X_5 to X_7 the causal effect on Y is consistently estimated with OLS, while we chose the detailed setup such that there is a detectable confounding bias on β_2^{OLS} and β_4^{OLS}. Thus, ideally, our local tests reject the null hypothesis for those two covariates but not for the rest.

For numerical results, we let the sample size grow from 10^2 to 10^6. For each sample size, we do 200 simulation runs. On the left-hand side of Figure 3.2, we show the average absolute z-statistics per predictor for the different sample sizes. For X_2 and X_4 we see the expected $\sqrt{n}$-growth. For the other variables, the empirical averages are close to the theoretical mean, which equals $\sqrt{2/\pi} \approx 0.8$, with a minimum of 0.70 and a maximum of 0.88. Further, we see that the confounding bias on the OLS parameter for X_4, which is a child of the hidden variable, is easier to detect than the bias onto

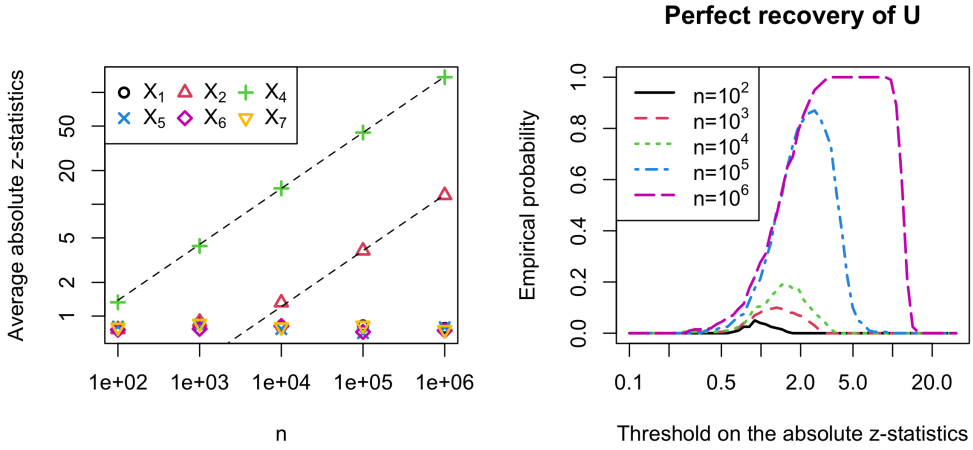


Figure 3.2: Simulation in a linear SEM corresponding to Figure 3.1. The results are based on 200 simulation runs. On the left: Average absolute z-statistics per covariate for different sample sizes. The dotted lines grow as $\sqrt{n}$ and are fit to match perfectly at $n = 10^5$. On the right: Empirical probability of perfectly recovering U (cf. (3.13)) for different sample sizes.

the parameter for X_2, which is a child's other parent. The multiplicity corrected p-value for X_4 is rejected at level $\alpha = 0.05$ in 91.5% of the cases for $n = 10^3$, while as the null hypothesis for X_2 is only rejected with a empirical probability of 3%. For X_2, it takes $n = 10^5$ samples to reject the local null hypothesis in 89% of the simulation runs.

Following Section 3.3.2, we should be able to perfectly recover the set U (cf. (3.13)) as $n \to \infty$ if we let the threshold on the absolute z-statistics grow at the right rate. Therefore, we plot on the right-hand side of Figure 3.2 the empirical probability of perfectly recovering U over a range of possible thresholds for the different sample sizes. For $n = 10^6$, we could achieve an empirical probability of 1. For $n = 10^5$ the optimum probability is 87%, while as for $n = 10^4$ it is only 19%.

Naturally, perfectly recovering U is a very ambitious goal for smaller sample sizes, and one might want to consider different objectives. In Figure 3.3, we plot two different performance metrics. On the left-hand side, we plot the empirical probability of not falsely including any variable in $\hat{U}$ against the average intersection size $\left|\hat{U} \cap U\right|$. The curve is parametrized implicitly by the threshold on the absolute z-statistics in order to reject the local null hypothesis for some variable. Thus, the graphic considers the question of how many variables in U can be recovered while keeping the probability of not falsely including any low. For a sample size of 10^5, we have an average intersection size of 3.97 allowing for a 10% probability of false inclusions. For 10^4, it is still 0.995. Thus, we can find (almost) one of the 4 variables in U on average. As we see in Figure 3.2, the bias on β_4^{OLS} is much easier

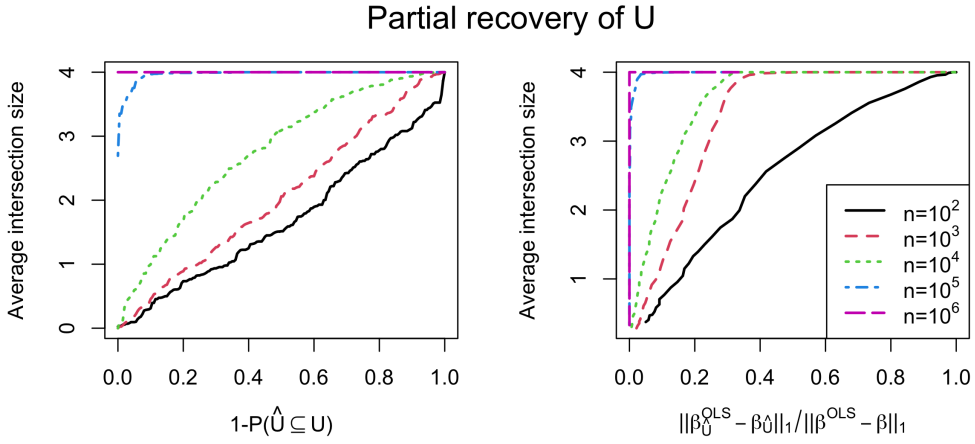


Figure 3.3: Simulation in a linear SEM corresponding to Figure 3.1. The results are based on 200 simulation runs. On the left: Probability of not falsely including a variable in $\hat{U}$ versus average intersection size $\left|\hat{U} \cap U\right|$ (cf. (3.15)). On the right: average remaining fraction of confounding signal versus average intersection size $\left|\hat{U} \cap U\right|$. It holds that $|U| = 4$. Both curves use the threshold on the absolute z-statistics as implicit curve parameter. Note that the legend applies to either plot.

to detect than the bias on β_2^{OLS}. Thus, keeping the probability of including X_2 in $\hat{U}$ low is still an ambitious task. Therefore, we analyse on the right-hand side of Figure 3.3 how many variables in U we can find while removing a certain amount of confounding signal. We define the remaining fraction as

$$\frac{\left\|\boldsymbol{\beta}_{\hat{U}}^{OLS} - \boldsymbol{\beta}_{\hat{U}}\right\|_1}{\left\|\boldsymbol{\beta}^{OLS} - \boldsymbol{\beta}\right\|_1},$$

i.e., how much of the difference $\boldsymbol{\beta}^{OLS} - \boldsymbol{\beta}$ persists in terms of ℓ_1 norm.

In this SEM, β_4^{OLS} caries 2/3 of the confounding signal, β_2^{OLS} only 1/3. Accepting 1/3 of remaining confounding signal, we receive an average intersection size of 3.885 for a sample size of 10^3. For 10^4, the average is 4. Thus, if we allow for false inclusion of X_2 we can almost perfectly retrieve all of U for sample size 10^3 already.

What if X includes descendants of Y? So far, we have only considered the case where $\mathbf{X}$ causally affects Y, but potentially, some of Y's parents are missing leading to a confounding effect. However, another possibility for $\boldsymbol{\beta}^{OLS}$ to not denote a causal effect is that there are descendants of Y amongst the predictors. The two different situations are depicted in Figure 3.4. The case with descendants in the set of predictors fits our theory from before if interpreted properly. If the model (3.4) for Y holds true using only the parents as predictors, Y can be naturally included in the assumed linear SEM for $\mathbf{X}$ in (3.16). Then, one can also think of $\mathcal{E}$ as an unobserved confounder. The Markov boundary of $\mathcal{E}$ with respect to the observed predictors is the same as the Markov boundary of Y. Of course, it holds $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, i.e., $\beta_j = 0 \; \forall j \notin \mathrm{PA}(Y)$. Using Theorem 3.9, we find

$$\beta_j^{HOLS} = \beta_j^{OLS} = \beta_j = 0 \; \forall j \in M \quad \text{that are not in the Markov boundary of } Y.$$

Thus, for all variables outside Y's Markov boundary, one can correctly detect that they have no causal effect onto Y ceteris paribus. The variables in the boundary, which includes all parents, are up to term cancellations all confounded. This can be detected under some conditions, as discussed in Section 3.3.2.

Figure 3.4: Left: SEM with a hidden confounder. Right: SEM with a descendant of Y.

3.4.3 Beyond linearity

We have mainly focused on linear models, i.e., the data is either generated by model (3.4) or model (3.9). Naturally, this assumption might be questionable in practice. Therefore, we provide some intuition about how HOLS might be applied in a more general setup. As we only detect misspecification of the OLS coefficient without identifying the type of misspecification, one should not try to over-interpret the effect of the regressors in $\hat{U}^c = \{1, \dots, p\} \setminus \hat{U}$ (cf. Section 3.3.2). However, the linear effect of the variables in $\hat{U}$ can always be interpreted to be well-specified, meaning that $\mathbb{E}[W_j|Z_j] = Z_j\beta_j^{OLS}$ or at least "sufficiently" well-specified such that no misspecification is detected in the data. Generally, we can write

$$\begin{aligned} Y &= \mathbf{X}^\top \boldsymbol{\beta}^{OLS} + f_{nonlinear}(\mathbf{X}) + \mathcal{E}, \text{ where } f_{nonlinear}(\mathbf{X}) = \mathbb{E}[Y|X] - \mathbf{X}^\top \boldsymbol{\beta}^{OLS}, \ \mathbb{E}[\mathcal{E}|\mathbf{X}] = 0. \\ W_j &= Z_j\beta_j^{OLS} + f_{nonlinear}(\mathbf{X}) + \mathcal{E}. \end{aligned}$$

Thus, well-specification of β_j^{OLS} implies $\mathbb{E}[f_{nonlinear}(\mathbf{X})|Z_j] = 0$, i.e., after linearly adjusting for $\mathbf{X}_{-j}$, X_j does not have any predictive power for $f_{nonlinear}(X)$. If it does not have any predictive power after linear adjustment, it would be a natural conclusion that it does not have predictive power after optimally adjusting for $\mathbf{X}_{-j}$ implying that $f_{nonlinear}(\mathbf{X})$ can be written as function of $\mathbf{X}_{-j}$ only. This would then imply

$$\mathbb{E}[Y|X_j = x_j + 1, \mathbf{X}_{-j} = \mathbf{x}_{-j}] - \mathbb{E}[Y|X_j = x_j, \mathbf{X}_{-j} = \mathbf{x}_{-j}] = \beta_j^{OLS} \quad \forall x_j, \mathbf{x}_{-j}.$$

Except for Gaussian data, such a linear relationship must be either causal or due to very pathological data setups. Excluding such unusual cancellations, the conclusion is that for $j \in U$ there must be a true linear causal effect from X_j to Y keeping the other predictors fixed, which can be consistently estimated using OLS. Of course, if there are no locally linear structures, it might well be that $U = \emptyset$ such that the local tests are not more informative than the global test. However, there is also nothing to be lost by exploiting this local view.

Note that the asymptotic results presented in Sections 3.3.1 and 3.3.2 hold for nonlinear data as well since they only assume model (3.11) - (3.12), which is the most general formulation.

3.5 Real data example

We analyse the flow cytometry dataset presented by Sachs et al. (2005). It contains cytometry measurements of 11 phosphorylated proteins and phospholipids. There is a "ground truth" on how these quantities affect each other, the so-called consensus network (Sachs et al., 2005). Data is available from various experimental conditions, some of which are interventional environments. The dataset has been further analysed in various projects, see, e.g., Mooij and Heskes (2013), Meinshausen et al.

(2016) and Taeb et al. (2023). Following these works, we consider data from 8 different environments, 7 of which are interventional. The sample size per environment ranges from 707 to 913.

In our analysis, we focus on the consensus network from Sachs et al. (2005). For each node, we go through all environments, fit a linear model using all its claimed parents as predictors, and assess the goodness of fit of the model using our HOLS check. In the consensus network, there is one bidirected edge between the variables PIP2 and PIP3. We include it as a parent for either direction. For each suggested edge, we also collect the p-values from the linear model fit in all environments, keeping only those where the edge passes the local HOLS check at level $\alpha = 5\%$ without multiplicity correction. We omit the multiplicity correction here to lower the tendency to falsely claim causal detection. In Table 3.1, we report the minimum p-value from OLS, over the environments where the HOLS check is passed, sorted by increasing p-values. Additionally, we show the number of environments in which the check is passed and out of these the number where the edge is significant at level $\alpha = 5\%$ in the respective linear model fit (with Bonferroni correction over all 8 environments and 17 edges, i.e., we require a p-value of at most 0.05/136). Note that there is one p-value 0 reported corresponding to a t-value of 174, which exceeds the precision that can be obtained with the standard R-function `lm`.

Edge	Passing HOLS	Significant in linear model	minimum p-value
RAF → MEK	3	2	0
PKA → Akt	3	3	1.5e-120
PKA → Erk	5	5	3.8e-69
PKC → JNK	3	3	5.9e-55
PIP2 → PIP3	1	1	6.5e-40
PIP3 → PLCg	5	1	1.4e-36
PKC → p38	1	1	7.1e-34
PIP3 → PIP2	1	1	9.6e-08
PLCg → PKC	6	0	0.016
PLCg → PIP2	1	0	0.027
PKC → RAF	8	0	0.046
PKC → PIP2	8	0	0.057
PKA → RAF	8	0	0.086
PKA → p38	8	0	0.12
PIP3 → Akt	8	0	0.2
PKA → JNK	8	0	0.21
MEK → Erk	8	0	0.42

Table 3.1: The working model is taken from the consensus network. The second column reports the number of environments in which the edge passes the HOLS check (among 8 possible ones). The third column additionally shows, in how many of these it is also significant in the respective linear model fit. The p-value is the minimum of the p-values from linear regression in environments, where the edge passes the HOLS check.

We see that the edge RAF $\to$ MEK is the most significant. Further, every edge of the consensus network passes the HOLS check in at least one environment. Frequently, we see that edges pass the HOLS check in certain environments without being significant in the linear model. Considering our discussion around linear SEMs, this could easily happen if the alleged predictor node is not actually in the Markov boundary of the response. In fact, there are seven edges that pass the HOLS check in every environment which are not significant based on the linear model fits. This is in agreement with Taeb et al. (2023), where none of them is reported.

As we cannot guarantee that the data follows a linear SEM as in (3.16), we shall not interpret the edges that do not pass the HOLS check to be subject to hidden confounding. However, the fact that we still find a decent number of suggested edges that pass the HOLS check, at least in some environments, leads to evidence that the assumption of some local unconfounded linear structures is not unrealistic, see also the discussion in Section 3.4.3.

We can also analyse our results in the light of invariant causal prediction, see, e.g., Peters et al. (2016), where one typically assumes that interventions do not change the underlying graph except for edges that point towards the node that is intervened on. This assumption is highly questionable in practice, and our findings, which vary a lot over different environments, indicate that the assumption is likely not fulfilled in the given setup.

3.6 Discussion

We have introduced the so-called HOLS check to assess the goodness of fit of linear causal models. It is based on the dependence between residuals and predictors in misspecified models, leading to non-vanishing higher moments. Besides checking whether the overall model might hold true, the method allows to detect a set of variable for which linear regression consistently estimates a true (unconfounded) causal effect for certain model classes.

We extend the HOLS method to high-dimensional datasets based on the idea of the debiased Lasso (Zhang and Zhang, 2014; van de Geer et al., 2014). This extension comes very naturally as our HOLS check involves nodewise regression just as the debiased Lasso.

Of particular interest are linear structural equation models, for which our method allows for very precise characterizations regarding which least squares parameters are causal effects. The result requires some non-Gaussianity. We complement our theory with a simulation study as well as a real data example.

A drawback of our method is that it does not distinguish whether a model is misspecified due to confounding or due to nonlinearities in the model. Therefore, an interesting follow-up direction would be to extend our methodology and theory from linear to nonlinear SEM using more flexible regression methods. This could allow to detect local causal structures in nonlinear settings as well.

Further simulation results as well as proofs and extended theory can be found in the supplemental material. Code scripts to reproduce the results presented in this paper are available here **https://github.com/cschultheiss/HOLS**.

3.A Simulation results

3.A.1 Global null

We create data that follows the model (3.4). We chose the sample size and dimensionality to be $n = 100$ and $p = 30$ and sample $\mathbf{X}$ as follows. Let $\Psi_1, \ldots, \Psi_p$ be i.i.d. random variables. Each of these follows a mixture distribution such that every copy comes from a $\mathcal{N}(0, 0.5)$ distribution with probability 2/3 or from a $\mathcal{N}(0, 2)$ distribution with probability 1/3. Thus, they are 0 mean and unit variance random variables. Then, set $X_1 = \Psi_1$ and

$$X_j = rX_{j-1} + \sqrt{1 - r^2}\Psi_j \quad (\forall j > 1)$$

This leads to a Toeplitz covariance structure $\mathbf{\Sigma}^{\mathbf{x}}$ with $\Sigma^{\mathbf{x}}_{ij} = r^{|i-j|}$, where we set $r = 0.6$. The coefficient vector $\boldsymbol{\beta}$ is 5-sparse, and the active predictors are $\{1, 5, 10, 15, 20\}$, each of which having a coefficient equal to 1. The random error $\mathcal{E}$ follows the same non-Gaussian distribution as the Ψ_j.

We run 200 simulations of this setup. For every simulation run, we calculate the p-value per predictor (without multiplicity adjustment) as well as the minimum of the multiplicity corrected p-values. Asymptotically, these p-values would be uniformly distributed as the model assumptions hold true. On the left-hand side of Figure 3.5, we analyse these p-values by looking at their empirical cumulative density function (ECDF). For p_j, the curve is combined over all the $p = 30$ covariates. Thus, it is based on 6000 p-values.

We see that even though the error does not have a Gaussian distribution, the p-values still are very close to being uniformly distributed. Furthermore, we see that the curve for the raw p-values is closer to the uniform distribution than the one for the minimum of the multiplicity corrected p-values. This is not surprising since the ECDF is based on more observations and as the CLT for multiple dimensions might take longer to kick in.

High-dimensional data We extend the simulation to a high-dimensional case. We reuse the setup with the only exception that $p = 200 > n$. Thus, we add an extra 170 predictors that do not actually have any influence on Y.

To calculate $\hat{\mathbf{z}}_j$, $\hat{\mathbf{w}}_j$, and $\hat{\sigma}$ we use the default implementation of the debiased Lasso, available in the R-package `hdi` (Dezeure et al., 2015). To get the estimate $\hat{\hat{\mathbf{z}}}^3_j$, we run a second level of nodewise regression with cross-validated $\tilde{\lambda}_j$. It shall be noted that all the estimated $\hat{\tilde{\boldsymbol{\gamma}}}_j = \mathbf{0}$. Thus, $\mathbf{x}_{-j}$ does not appear to contain strong enough information on $\mathbf{z}^3_j$, at least for the given sample size.

Again, we look at all the obtained raw p-values and plot the ECDF on the right-hand side of Figure 3.5. We see some deviations from the uniform distribution. Especially, very low p-values become more unlikely. Thus, the procedure is a bit too conservative. This is emphasized by the

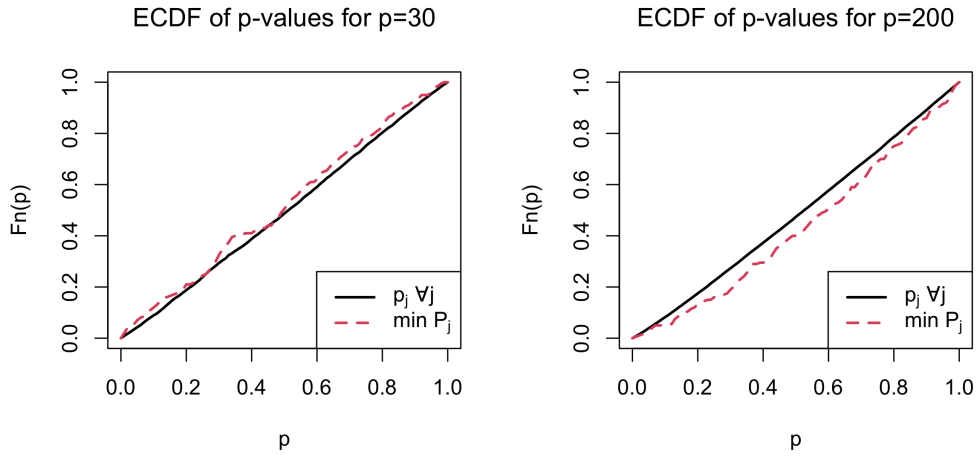


Figure 3.5: Simulation under the global null. The results are based on 200 simulation runs. On the left: ECDF of p-values for low-dimensional data p_j (unadjusted as in Step 9 of Algorithm 3.1) over all $p = 30$ predictors and of $\min_j P_j$ (multiplicity corrected as in Step 10 of Algorithm 3.1). On the right: ECDF of p-values for high-dimensional data p_j (unadjusted as in Step 10 of Algorithm 3.2) over all $p = 200$ predictors and of $\min_j P_j$ (multiplicity corrected as in Step 11 of Algorithm 3.2.

ECDF for the minimum of the multiplicity corrected p-values as this minimum is affected by the distribution of very low p-values. The issue might be related to σ being overestimated: the empirical average of $\hat{\sigma}^2$ is 1.35 and $\hat{\sigma} > \sigma = 1$ occurred in 88.5% of the cases. However, when replacing the estimate with the true σ the p-values become too liberal.

In summary, after increasing the number of predictors but keeping the sample size the same, the results deviate a bit more from the optimal distribution. Though, the behaviour is still fairly close to what one would aim for, supporting the benefit of our method.

3.A.2 Missing variable in a linear SEM

For the sake of comparison with the results under the global null, we provide here an additional analysis of the simulation example in Section 3.4.2. Namely, we show in Figure 3.6 the empirical cumulative distribution function of the p-values obtained for different predictors. The predictors X_2 and X_4, for which the local null hypothesis should be rejected, are depicted separately, while as the rest is grouped together. Note that for X_4 sample sizes of more than 10^4 are not in the plot anymore as they look just the same.

We see that the distribution of the p-values for the covariates outside the hidden variable's Markov boundary are close to the desired uniform distribution even for low sample sizes. Furthermore, in accordance with the z-statistics shown in Section 3.4.2, the confounding bias on β_4^{OLS} is much easier to detect than the bias on β_2^{OLS}.

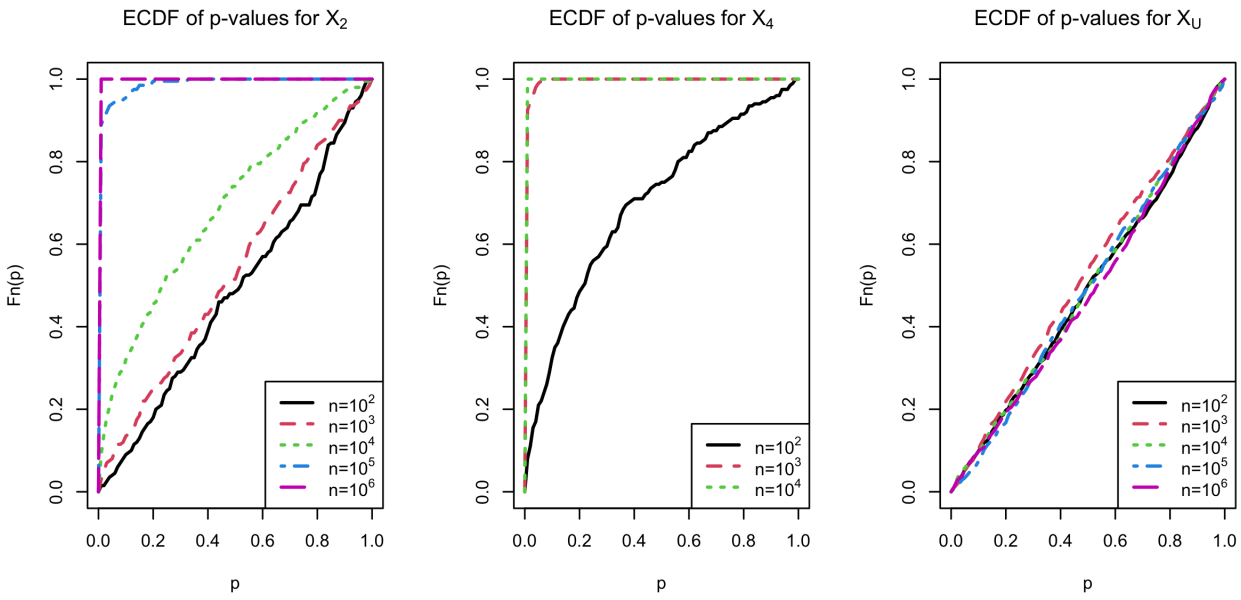


Figure 3.6: Simulation in a linear SEM corresponding to Figure 3.1. The results are based on 200 simulation runs. Depicted is the ECDF of the p-values for different predictors (unadjusted p_j as in Step 9 of Algorithm 3.1).

3.A.3 High-dimensional data: missing variable in a linear SEM

We want to assess how well our method for high-dimensional data can detect deviations from the null hypothesis. We create data from the linear SEM depicted in Figure 3.7, where all predictors but

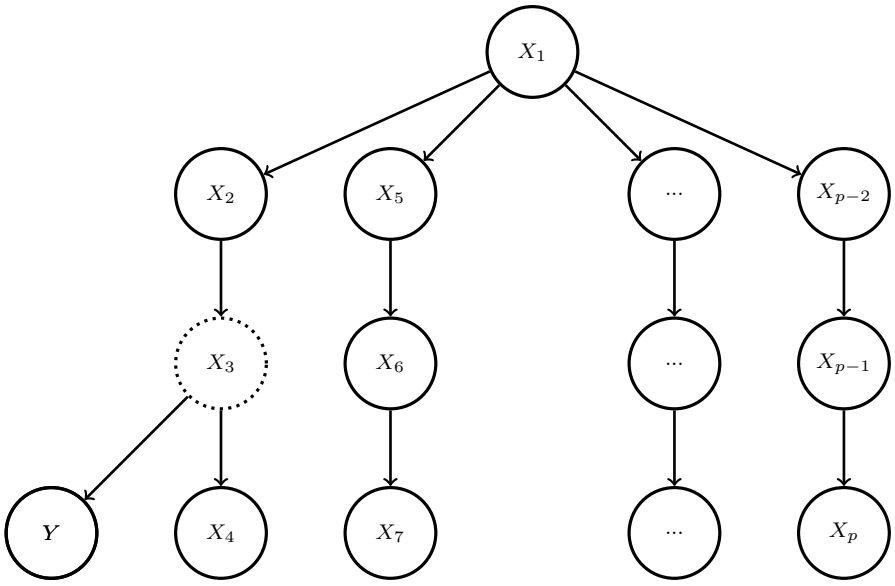


Figure 3.7: Linear SEM used for the high-dimensional simulation.

X_3 are observed. According to our theoretical results, $\beta_j^{OLS} = \beta_j^{HOLS} = \beta_j \ \forall j \notin \{2, 4\}$. We set the number of variables to $p = 1.5n + 1$. Thus, there are $0.5n$ of these blocks. We consider $n = 10^2$ and $n = 10^3$. For a comparison with low-dimensional HOLS, we also assess the performance using just $p = 13$ predictors, i.e., four blocks. We let all Ψ_j follow a centered uniform distribution and set $Y = X_3$. To execute the high-dimensional HOLS test, we proceed as in Section 3.A.1.

In Figure 3.8, we show the empirical cumulative distribution function of the p-values obtained over 200 simulation runs. We look at the p-values for X_2 and X_4 separately and for all other variables combined. The latter should roughly follow a uniform distribution. Similar to our results in Sections 3.4.2 and 3.A.2, we see that it is much easier to reject the null hypothesis for the hidden variable's child X_4 than for the child's other parent (with respect to the observed covariates) X_2. With a sample size of 10^3, no p-value p_4 larger than $1.4 * 10^{-4}$ was obtained. Finally, we see that for the other covariates the obtained p-values are indeed close to being uniformly distributed.

In Figure 3.9, we consider the same statistics for the low-dimensional HOLS test applied to the same data but with just the first 12 observed covariates. We note that for this data the distribution of the p-values for X_2 and X_4 is more distinct from the uniform distribution in the high-dimensional setup than in the low-dimensional setup. Other than that, the conclusion regarding the algorithm's performance remains the same.

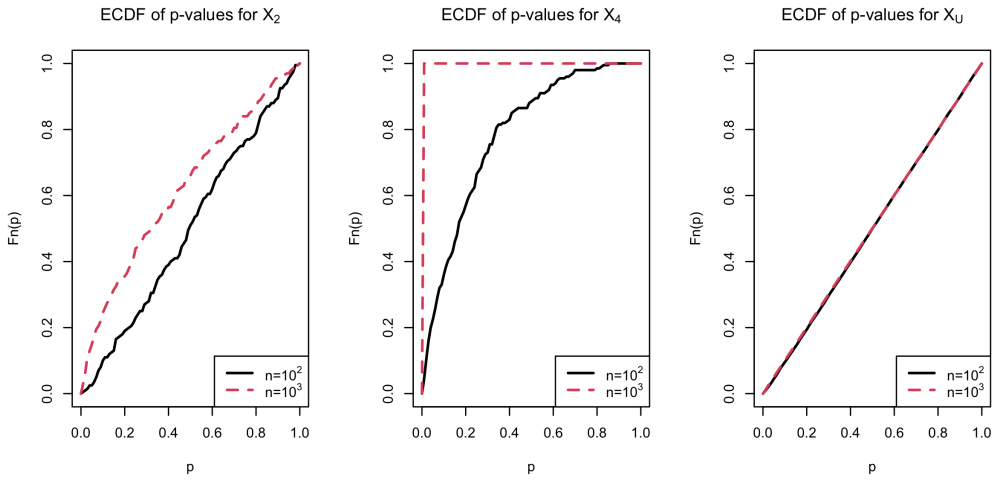


Figure 3.8: Simulation for a missing variable in a linear SEM corresponding to Figure 3.7 for high-dimensional data. The results are based on 200 simulation runs. Depicted is the ECDF of the p-values for different predictors (unadjusted p_j as in Step 10 of Algorithm 3.2).

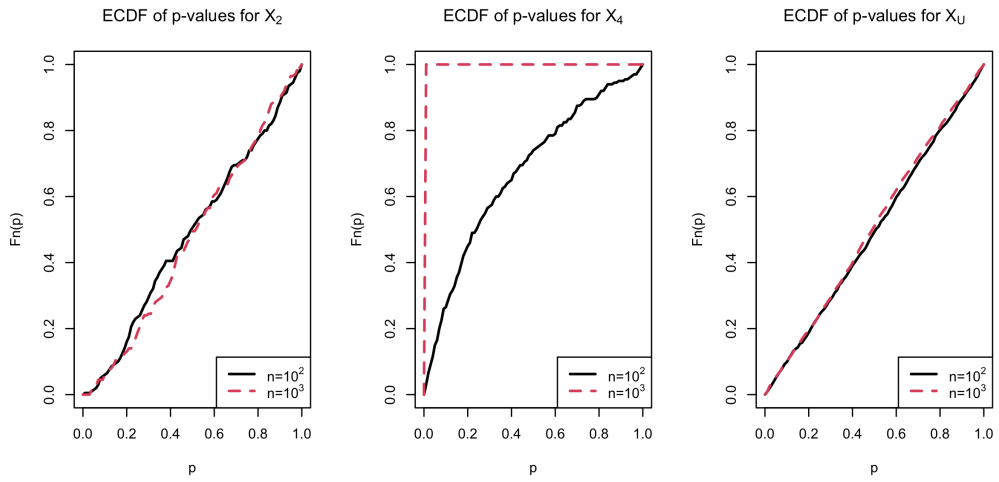


Figure 3.9: Simulation for a missing variable in a linear SEM corresponding to Figure 3.7 for low-dimensional data. The results are based on 200 simulation runs. Depicted is the ECDF of the p-values for different predictors (unadjusted p_j as in Step 9 of Algorithm 3.1).

3.A.4 Confounding onto block-independent $\mathcal{E}_{\mathbf{X}}$

To simulate block-independent data, we make use of the Boston housing data, available in the R-package `MASS` (Venables and Ripley, 2002). We use all variables but the variable `medv`, which is typically the response variable for regression. Then, we create two independent bootstrap samples and concatenate those such that we have two independent blocks forming the matrix $\boldsymbol{\epsilon}_{\mathbf{X}}$. There are 13 covariates per block. Before this bootstrap sampling, we make all variables have 0 mean and unit variance. We let H be a standard normal random variable and set $X_1 = \mathcal{E}_{X_1} + H$ and $X_7 = \mathcal{E}_{X_7} - H$. For the remaining variables, we let $X_j = \mathcal{E}_{X_j}$. Finally, we set $Y = H$ for simplicity. Thus, the first block is confounded with Y, but the second is not. As the covariance $\boldsymbol{\Sigma}^{\mathcal{E}_{\mathbf{X}}}$ is defined by the empirical correlation of the Boston housing data, there are no vanishing entries in each of the blocks. Thus, none of the covariates X_1 to X_{13} fulfils (3.14), and there is a confounding bias on each of the OLS parameters.

We vary the sample size from 10^2 to 10^6, doing 200 simulation runs for each sample size. Thus, it is a "m out of n" bootstrap, where m can be smaller than n (for $m = 10^2$) or larger (for the rest). In the remainder, we call the bootstrap sample size n to keep the notation consistent and as the size of the real Boston housing data is not of primary interest. On the left-hand side of Figure 3.10, we plot the average absolute z-statistics for a representative subset of the predictors. Notably, it is the same four predictors once from the first block and once from the second. As expected, this average grows as $\sqrt{n}$ for variables in the confounded block, while it stays approximately constant for variables from the independent block. Further, we see that the two variables X_1 and X_7, which are directly confounded,

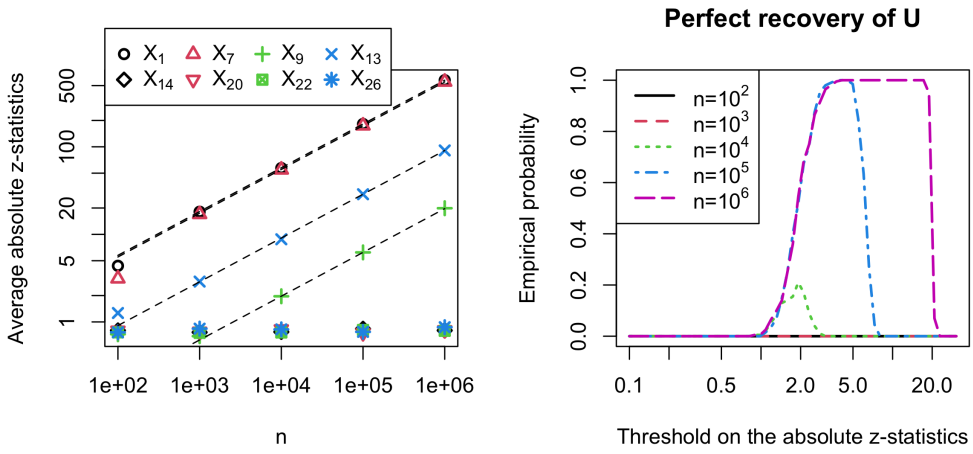


Figure 3.10: Simulation for confounding onto block independent $\mathcal{E}_{\mathbf{X}}$. The results are based on 200 simulation runs. On the left: Average absolute z-statistics per covariate for different sample sizes. The dotted lines grow as $\sqrt{n}$ and are fit to match perfectly at $n = 10^5$. On the right: Empirical probability (over 200 simulation runs) of perfectly recovering U (cf. (3.13)) for different sample sizes.

are the easiest to detect as such. For only 10^2 samples, the multiplicity corrected p-values for X_1 and X_7 lead to a rejection of the local null hypothesis in 78.5% respectively 46% of the simulation runs at level $\alpha = 0.05$. For some of the other variables in the confounded block, it takes many more samples to reliably reject the local null hypothesis. For X_9, the local null hypothesis is only rejected in 3% of the cases for $n = 10^4$ and only from $n = 10^5$ it is always rejected.

On the right-hand side, we show the empirical probability of perfectly recovering U, i.e., rejecting the null hypothesis for all variables from the first block but not rejecting it for any variable from the second block. We see that for $n = 10^5$ we are able to achieve this recovery with an empirical probability of 1, and, for $n = 10^6$, it is even possible for a larger range of thresholds. Comparing the two curves for $n = 10^5$ and $n = 10^6$, we see that they initially look very similar. This is as one would expect as the initial increase of the curve corresponds to reducing the type I error, which is independent of the sample size, assuming the CLT has kicked in sufficiently. The decrease of the curve depends on the z-statistics for the confounded variables, which we know to increase as $\sqrt{n}$. Thus, this decrease will appear later the larger n gets.

We show the empirical cumulative distribution function of the obtained p-values in Figure 3.11. X_1 and X_9 are considered separately while as all p-values for the variables from the second block are grouped together. For the latter, the ECDF is close to the desired uniform distribution for every sample size. In accordance with the average z-statistics, the p-values for the directly confounded variable X_1 are more extreme than those for X_9.

In Figure 3.12, we analyse the partial recovery of U as in Section 3.4.2. We see that for a sample

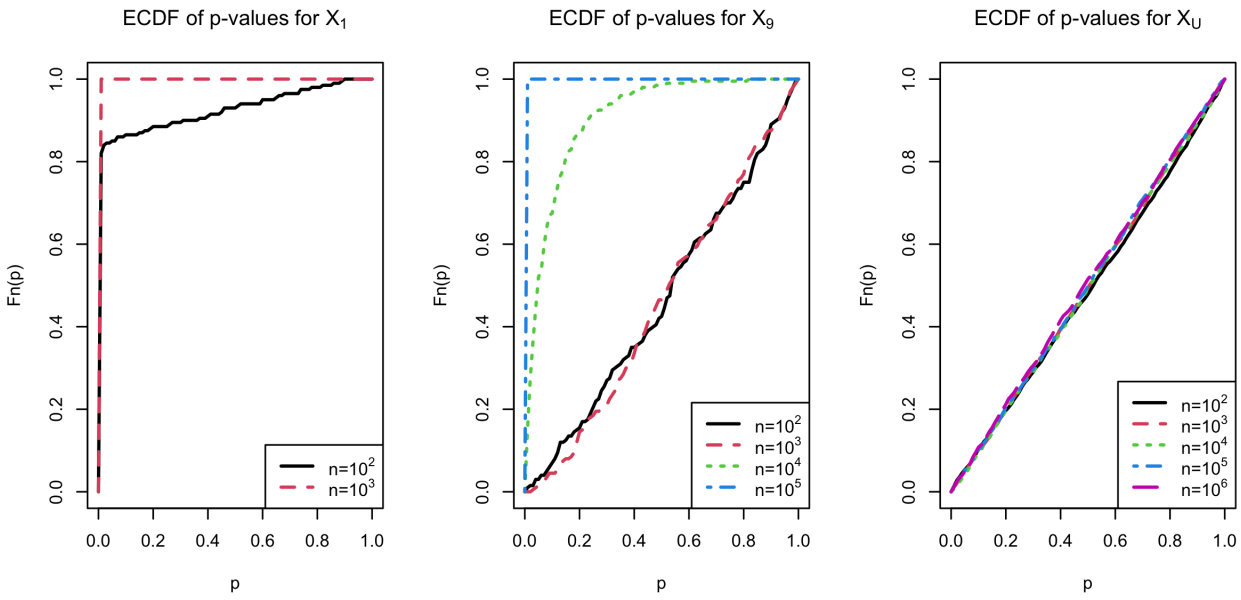


Figure 3.11: Simulation for confounding onto block independent $\mathcal{E}_{\mathbf{X}}$. The results are based on 200 simulation runs. Depicted is the ECDF of the p-values for different predictors (unadjusted p_j as in Step 9 of Algorithm 3.1).

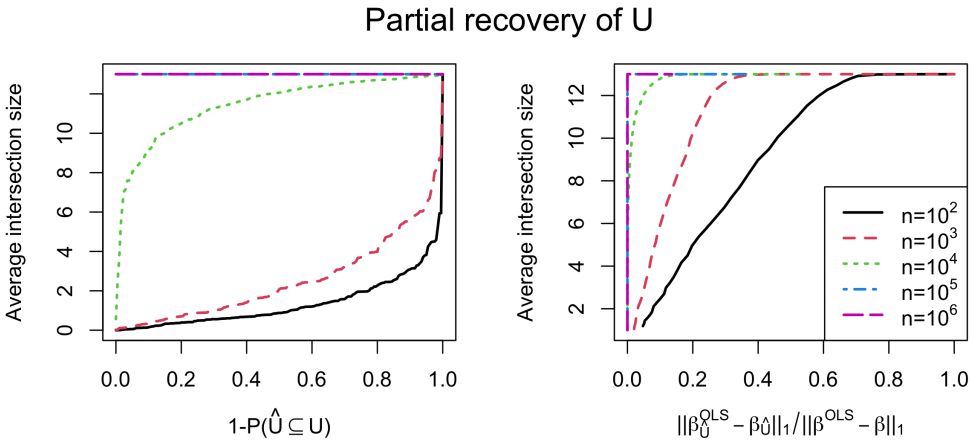


Figure 3.12: Simulation for confounding onto block independent $\mathcal{E}_{\mathbf{X}}$. The results are based on 200 simulation runs. On the left: Probability of not falsely including a variable in $\hat{U}$ versus average intersection size $\left|\hat{U} \cap U\right|$ (cf. (3.15)). On the right: average remaining fraction of confounding signal versus average intersection size $\left|\hat{U} \cap U\right|$. It holds that $|U| = 13$. Both curves use the threshold on the absolute z-statistics as implicit curve parameter. Note that the legend applies to either plot.

size of 10^4, for which perfect recovery is hardly achievable, we receive an average intersection size of 9.175 (out of 13) allowing for 10% probability of false inclusion. For lower sample sizes, there is not much that can be found under this constraint. In this setup, there is a confounding bias onto 13 OLS parameters with varying signal strength $\left|\beta_j^{OLS} - \beta_j\right|$. The two directly confounded variables amount to 45.254% of the confounding signal. Thus, a remaining fraction of 54.746% appears to be particularly achievable. We see that we can do even better than that. Namely, for a sample size of 10^3, we can get an empirical probability of 1 of including all variables of U in $\hat{U}$ allowing for an average of 44.775% of the confounding signal. For 10^4, we can go down to a remaining fraction of 17.597%.

3.B Proofs

3.B.1 Proof of Theorem 3.1

Note that model (3.4) and (A3.2) imply that (B3.1) - (B3.3) hold for all j, i.e., $U' = \{1, \dots, p\}$. Thus, we receive Theorem 3.1 for free by proving Theorem 3.5 and we receive Corollary 3.1 for free by proving Corollary 3.2.

3.B.2 Proof of Theorem 3.3

Note first that Assumptions (A3.1) and (A3.2) imply

$$\begin{aligned} \frac{1}{n}\mathbf{x}_{-j}^\top \mathbf{x}_{-j} \xrightarrow{\mathbb{P}} \mathbf{\Sigma}^{\mathbf{X}}_{-j,-j} \implies & n\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)^{-1} \xrightarrow{\mathbb{P}} \big(\mathbf{\Sigma}^{\mathbf{X}}_{-j,-j}\big)^{-1} \\ \implies & \left\| n\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)^{-1} \right\| \xrightarrow{\mathbb{P}} \left\| \big(\mathbf{\Sigma}^{\mathbf{X}}_{-j,-j}\big)^{-1} \right\| = \mathcal{O}(1), \end{aligned} \tag{3.18}$$

where we use invertibility and the continuous mapping theorem. In several occasions, we use bounds on multiplication with the projection matrix $\boldsymbol{P}_{-j}$, e.g.,

$$\begin{aligned} \left|\mathbf{z}_j^\top \boldsymbol{P}_{-j} \mathbf{w}_j\right| &= \left|\mathbf{z}_j^\top \mathbf{x}_{-j}\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)^{-1} \mathbf{x}_{-j}^\top \mathbf{w}_j\right| \leq \left\|\mathbf{z}_j^\top \mathbf{x}_{-j}\right\|_2 \left\|\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)^{-1}\right\|_2 \left\|\mathbf{x}_{-j}^\top \mathbf{w}_j\right\|_2 \\ &\leq \left\|\mathbf{z}_j^\top \mathbf{x}_{-j}\right\|_1 \left\|\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)^{-1}\right\|_2 \left\|\mathbf{x}_{-j}^\top \mathbf{w}_j\right\|_1 = \sum_{k\neq j}\left|\mathbf{z}_j^\top \mathbf{x}_k\right| \left\|\Big(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\Big)^{-1}\right\|_2 \sum_{k\neq j}\left|\mathbf{x}_k^\top \mathbf{w}_j\right| \\ &= \mathcal{O}_p(\sqrt{n})\mathcal{O}_p\left(\frac{1}{n}\right)o_p(n). \end{aligned} \tag{3.19}$$

For the last equality, we used Chebyshev's inequality, (3.18), and the LLN together with $\mathbb{E}[Z_j X_k] = 0$, $\mathbb{E}\Big[(Z_j X_k)^2\Big] < \infty$ and $\mathbb{E}[X_k W_j] = 0$.

Theorem 3.3 consists of three parts. Consider $\hat{\beta}_j^{OLS}$.

$$\begin{aligned}
\frac{1}{n}\left|\hat{\mathbf{z}}_j^\top \hat{\mathbf{w}}_j - \mathbf{z}_j^\top \mathbf{w}_j\right| &= \frac{1}{n}\left|\mathbf{z}_j^\top \boldsymbol{P}_{-j}^\perp \boldsymbol{P}_{-j}^\perp \mathbf{w}_j - \mathbf{z}_j^\top \mathbf{w}_j\right| = \frac{1}{n}\left|\mathbf{z}_j^\top \boldsymbol{P}_{-j} \mathbf{w}_j\right| \\
&= \frac{1}{n}\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\left(\frac{1}{n}\right)\mathcal{o}_p(n) = \mathcal{o}_p\left(\frac{1}{\sqrt{n}}\right). \text{ Thus,} \\
\frac{1}{n}\hat{\mathbf{z}}_j^\top \hat{\mathbf{w}}_j &= \frac{1}{n}\mathbf{z}_j^\top \mathbf{w}_j + \mathcal{o}_p\left(\frac{1}{\sqrt{n}}\right) = \mathbb{E}[Z_j W_j] + \mathcal{o}_p(1) + \mathcal{o}_p\left(\frac{1}{\sqrt{n}}\right) = \mathbb{E}[Z_j W_j] + \mathcal{o}_p(1). \\
\frac{1}{n}\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j &= \mathbb{E}\left[Z_j^2\right] + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) \text{ follows analogously such that} \\
\hat{\beta}_j^{OLS} &= \frac{\mathbb{E}[Z_j W_j]}{\mathbb{E}\left[Z_j^2\right]} + \mathcal{o}_p(1).
\end{aligned}$$

For $\hat{\beta}_j^{HOLS}$, we first consider some intermediate results.

$$\begin{aligned}
\left\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right\|_2 &= \left\|\left(\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\right)^{-1} \mathbf{x}_{-j}^\top \mathbf{z}_j\right\|_2 \leq \left\|\left(\frac{1}{n}\mathbf{x}_{-j}^\top \mathbf{x}_{-j}\right)^{-1}\right\|_2 \left\|\frac{1}{n}\mathbf{x}_{-j}^\top \mathbf{z}_j\right\|_2 \\
&= \mathcal{O}_p(1)\mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) \text{ such that} \\
\left\|\hat{\mathbf{z}}_j - \mathbf{z}_j\right\|_\infty &= \left\|\mathbf{x}_{-j}\left(\boldsymbol{\gamma}_j - \hat{\boldsymbol{\gamma}}_j\right)\right\|_\infty \leq \left\|\mathbf{x}_{-j}\right\|_\infty \left\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right\|_1 \leq \left\|\mathbf{x}_{-j}\right\|_\infty \sqrt{p}\left\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right\|_2 \\
&= \mathcal{O}_p(K)\mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) = \mathcal{O}_p\left(\frac{K}{\sqrt{n}}\right),
\end{aligned}$$

using fixed p. Note that we denote the bound on $\left\|\mathbf{x}_{-j}\right\|_\infty$ by K. (A3.2) induces a worst-case bound of $K = n^{1/6}$. This could be heavily improved for certain assumptions on the distribution of $\mathbf{X}$, e.g., $K = \sqrt{\log(n)}$ for Gaussian data. To keep things more general, we will use generic K in the following. Further,

$$\left\|\hat{\mathbf{z}}_j - \mathbf{z}_j\right\|_2^2 = \left\|\boldsymbol{P}_{-j}^\perp \mathbf{z}_j - \mathbf{z}_j\right\|_2^2 = \left\|\boldsymbol{P}_{-j}\mathbf{z}_j\right\|_2^2 = \mathcal{O}_p(1) \quad \text{and analogously} \quad \left\|\hat{\mathbf{w}}_j - \mathbf{w}_j\right\|_2^2 = \mathcal{o}_p(n).$$

We invoke the following identity

$$\left(a^3 - b^3\right) = (a-b)^3 - 3a(a-b)^2 + 3a^2(a-b)$$

to find

$$\begin{aligned}\left\|\mathbf{z}_j^3-\hat{\mathbf{z}}_j^3\right\|_2 &\leq \left\|(\mathbf{z}_j-\hat{\mathbf{z}}_j)^3\right\|_2+3\left\|\mathbf{z}_j\odot(\mathbf{z}_j-\hat{\mathbf{z}}_j)^2\right\|_2+3\left\|\mathbf{z}_j^2\odot(\mathbf{z}_j-\hat{\mathbf{z}}_j)\right\|_2\\ &\leq \left\|(\mathbf{z}_j-\hat{\mathbf{z}}_j)^2\right\|_\infty\|\mathbf{z}_j-\hat{\mathbf{z}}_j\|_2+3\|\mathbf{z}_j\|_\infty\|\mathbf{z}_j-\hat{\mathbf{z}}_j\|_\infty\|\mathbf{z}_j-\hat{\mathbf{z}}_j\|_2+3\left\|\mathbf{z}_j^2\right\|_\infty\|(\mathbf{z}_j-\hat{\mathbf{z}}_j)\|_2 \quad (3.20)\\ &= \mathcal{O}_p\Big(\frac{K^2}{n}\Big)+\mathcal{O}_p\Big(\frac{K^2}{\sqrt{n}}\Big)+\mathcal{O}_p\big(K^2\big)=\mathcal{O}_p\big(K^2\big).\end{aligned}$$

With this at hand, we find

$$\begin{aligned}&\frac{1}{n}\Big|\big(\mathbf{z}_j^3\big)^\top\mathbf{w}_j-\big(\hat{\mathbf{z}}_j^3\big)^\top\hat{\mathbf{w}}_j\Big|\\ =&\frac{1}{n}\Big|\big(\mathbf{z}_j^3-\hat{\mathbf{z}}_j^3\big)^\top\mathbf{w}_j+\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)^\top(\mathbf{w}_j-\hat{\mathbf{w}}_j)+\big(\mathbf{z}_j^3\big)^\top(\mathbf{w}_j-\hat{\mathbf{w}}_j)\Big|\\ \leq&\frac{1}{n}\Big(\Big|\big(\mathbf{z}_j^3-\hat{\mathbf{z}}_j^3\big)^\top\mathbf{w}_j\Big|+\Big|\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)^\top(\mathbf{w}_j-\hat{\mathbf{w}}_j)\Big|+\Big|\big(\mathbf{z}_j^3\big)^\top(\mathbf{w}_j-\hat{\mathbf{w}}_j)\Big|\Big)\\ \leq&\frac{1}{n}\Big(\big\|\big(\mathbf{z}_j^3-\hat{\mathbf{z}}_j^3\big)\big\|_2\|\mathbf{w}_j\|_2+\big\|\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)\big\|_2\|(\mathbf{w}_j-\hat{\mathbf{w}}_j)\|_2+\big\|\big(\mathbf{z}_j^3\big)\big\|_2\|(\mathbf{w}_j-\hat{\mathbf{w}}_j)\|_2\Big)\\ =&\frac{1}{n}\big(\mathcal{O}_p\big(K^2\big)\mathcal{O}_p\big(\sqrt{n}\big)+\mathcal{O}_p\big(K^2\big)o_p\big(\sqrt{n}\big)+\mathcal{O}_p\big(\sqrt{n}\big)o_p\big(\sqrt{n}\big)\big)=o_p(1).\text{ Thus,}\\ \frac{1}{n}\big(\hat{\mathbf{z}}_j^3\big)^\top\hat{\mathbf{w}}_j=&\frac{1}{n}\big(\mathbf{z}_j^3\big)^\top\mathbf{w}_j+o_p(1)=\mathbb{E}\big[Z_j^3W_j\big]+o_p(1).\\ \frac{1}{n}\big(\hat{\mathbf{z}}_j^3\big)^\top\hat{\mathbf{z}}_j=&\mathbb{E}\big[Z_j^4\big]+\mathcal{O}_p\Big(\frac{K^2}{\sqrt{n}}\Big)\text{ follows analogously such that}\\ \hat{\beta}_j^{HOLS}=&\frac{\mathbb{E}\Big[Z_j^3W_j\Big]}{\mathbb{E}\Big[Z_j^4\Big]}=\frac{\mathbb{E}\Big[Z_j^3W_j\Big]}{\mathbb{E}\Big[Z_j^4\Big]}+o_p(1).\end{aligned}$$

The last part of Theorem 3.3 considers the variance estimate (cf. (3.8)) and is implied by the following Lemma which is a more precise statement.

Lemma 3.1. Assume that the data follows the model (3.11) and that (A3.1) - (A3.2) hold. Then,

$$\widehat{\mathrm{Var}}\Big(\sqrt{n}\Big(\hat{\beta}_j^{HOLS}-\hat{\beta}_j^{OLS}\Big)\Big)\xrightarrow{\mathbb{P}}\sigma_{\tilde{\mathcal{E}}}^2\left(\frac{\mathbb{E}\Big[\Big(\tilde{Z}_j^3\Big)^2\Big]}{\mathbb{E}\Big[Z_j^4\Big]^2}-\frac{1}{\mathbb{E}\Big[Z_j^2\Big]}\right)\ \forall j.$$

Note that we defined

$$\widehat{\mathrm{Var}}\Big(\sqrt{n}\Big(\hat{\beta}_j^{HOLS}-\hat{\beta}_j^{OLS}\Big)\Big):=n\widehat{\mathrm{Var}}\Big(\Big(\hat{\beta}_j^{HOLS}-\hat{\beta}_j^{OLS}\Big)\Big).$$

3.B.2.1 Proof of Lemma 3.1

For $\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j$ and $\left(\hat{\mathbf{z}}_j^2\right)^\top (\hat{\mathbf{z}}_j)^2 = \left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{z}}_j$, we have established convergence already. It remains to look at the other terms in (3.8), i.e., $\left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\hat{\mathbf{z}}_j^3\right)$ and $\hat{\sigma}^2$. We find

$$\begin{aligned}
&\frac{1}{n}\left|\left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\hat{\mathbf{z}}_j^3\right) - \left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\mathbf{z}_j^3\right)\right| = \frac{1}{n}\left|\left(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right) + 2\left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right)\right| \\
\leq \frac{1}{n}\left\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right\|_2^2 + \frac{2}{n}\left\|\mathbf{z}_j^3\right\|_2 \left\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right\|_2 &= \mathcal{O}_p\left(\frac{K^4}{n}\right) + \mathcal{O}_p\left(\frac{K^2}{\sqrt{n}}\right) = \mathcal{O}_p\left(\frac{K^2}{\sqrt{n}}\right), \text{ and} \\
\frac{1}{n}\left|\left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\mathbf{z}_j^3\right) - \left(\tilde{\mathbf{z}}_j^3\right)^\top \left(\tilde{\mathbf{z}}_j^3\right)\right| = \frac{1}{n}\left|\left(\tilde{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\tilde{\mathbf{z}}_j^3\right) - \left(\tilde{\mathbf{z}}_j^3\right)^\top \left(\tilde{\mathbf{z}}_j^3\right)\right| &= \frac{1}{n}\left|\left(\tilde{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j} \left(\tilde{\mathbf{z}}_j^3\right)\right| \\
&= \mathcal{O}_p(1) \text{ such that} \\
\frac{1}{n}\left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \left(\hat{\mathbf{z}}_j^3\right) = \frac{1}{n}\left(\tilde{\mathbf{z}}_j^3\right)^\top \left(\tilde{\mathbf{z}}_j^3\right) + \mathcal{O}_p(1) &= \mathbb{E}\left[\left(\tilde{Z}_j^3\right)^2\right] + \mathcal{O}_p(1).
\end{aligned}$$

This ensures convergence of the per variable error scaling. It remains to estimate the variance of $\tilde{\mathcal{E}}$. Although the error is now only uncorrelated but not independent from $\mathbf{X}$ (cf. (3.11)), the variance can still be estimated consistently using the standard formula. Let

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}^{OLS} = \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}},$$

which is used for variance estimation. We find

$$\begin{aligned}
\frac{1}{n-p}\left|\tilde{\boldsymbol{\epsilon}}^\top \tilde{\boldsymbol{\epsilon}} - \hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}\right| = \frac{1}{n-p}\left|\tilde{\boldsymbol{\epsilon}}^\top \boldsymbol{P}_{-j} \tilde{\boldsymbol{\epsilon}}\right| = \mathcal{O}_p\left(\frac{1}{n}\right)\mathcal{O}_p(n)\mathcal{O}_p\left(\frac{1}{n}\right)\mathcal{O}_p(n) = \mathcal{O}_p(1) \text{ such that} \\
\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\epsilon}}\|_2^2}{n-p} = \frac{\|\tilde{\boldsymbol{\epsilon}}\|_2^2}{n-p} + \mathcal{O}_p(1) = \mathbb{E}\left[\tilde{\mathcal{E}}^2\right] + \mathcal{O}_p(1).
\end{aligned}$$

3.B.3 Proof of Theorem 3.4

We provide a supporting Lemma.

Lemma 3.2. Assume that the data follows the model (3.11) and that (A3.1) - (A3.3) hold . Let j be some covariate with $\beta_j^{OLS} = \beta_j^{HOLS}$ for which (B3.1) and (A3.4) hold. Then,

$$\sqrt{n}\left(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\right) \xrightarrow{\mathbb{D}} \mathcal{N}\left(0, \mathrm{Var}\left(\frac{\tilde{Z}_j^3 \tilde{\mathcal{E}}}{\mathbb{E}\left[Z_j^4\right]} - \frac{Z_j \tilde{\mathcal{E}}}{\mathbb{E}\left[Z_j^2\right]}\right)\right).$$

If (B3.2) and (B3.3) hold as well for j, this can be refined as

$$\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big) \xrightarrow{\mathbb{D}} \mathcal{N}\left(0, \sigma_{\tilde{\mathcal{E}}}^2\left(\frac{\mathbb{E}\Big[\Big(\tilde{Z}_j^3\Big)^2\Big]}{\mathbb{E}\Big[Z_j^4\Big]^2} - \frac{1}{\mathbb{E}\Big[Z_j^2\Big]}\right)\right).$$

Note that (A3.4) is implied by (B3.3). Theorem 3.4 follows from Lemmata 3.1 and 3.2, applying Slutsky's theorem. Thus, it remains to prove Lemma 3.2.

3.B.3.1 Proof of Lemma 3.2

We look at the scaled estimates $\sqrt{n}\hat{\beta}_j^{OLS}$ and $\sqrt{n}\hat{\beta}_j^{HOLS}$ for some variable with $\beta_j^{OLS} = \beta_j^{HOLS}$ fulfilling (B3.1) and (A3.4). Note that since we assume (A3.3), we can sharpen $\left|\mathbf{x}_k^\top \mathbf{w}_j\right| = \mathcal{O}_p(\sqrt{n})$ instead of just $\mathcal{O}_p(n)$.

$$\begin{aligned}
\sqrt{n}\hat{\beta}_j^{OLS} &= \frac{\sqrt{n}\frac{1}{n}\hat{\mathbf{z}}_j^\top \hat{\mathbf{w}}_j}{\frac{1}{n}\hat{\mathbf{z}}_j^\top \hat{\mathbf{z}}_j} = \frac{\sqrt{n}\frac{1}{n}\mathbf{z}_j^\top \mathbf{w}_j + \mathcal{O}_p(1/\sqrt{n})}{\frac{1}{n}\mathbf{z}_j^\top \mathbf{z}_j + \mathcal{O}_p(1/n)} = \frac{\sqrt{n}\frac{1}{n}\mathbf{z}_j^\top \mathbf{w}_j}{\frac{1}{n}\mathbf{z}_j^\top \mathbf{z}_j} + \mathcal{O}_p\big(1/\sqrt{n}\big) \\
&= \sqrt{n}\beta_j^{OLS} + \frac{\sqrt{n}\frac{1}{n}\mathbf{z}_j^\top \tilde{\boldsymbol{\epsilon}}}{\frac{1}{n}\mathbf{z}_j^\top \mathbf{z}_j} + \mathcal{O}_p\big(1/\sqrt{n}\big) = \sqrt{n}\beta_j^{OLS} + \frac{\sqrt{n}\frac{1}{n}\mathbf{z}_j^\top \tilde{\boldsymbol{\epsilon}}}{\mathbb{E}\Big[Z_j^2\Big] + \mathcal{O}_p(1/\sqrt{n})} + \mathcal{O}_p\big(1/\sqrt{n}\big) \\
&= \sqrt{n}\beta_j^{OLS} + \frac{\sqrt{n}\frac{1}{n}\mathbf{z}_j^\top \tilde{\boldsymbol{\epsilon}}}{\mathbb{E}\Big[Z_j^2\Big]} + \mathcal{O}_p\big(1/\sqrt{n}\big),
\end{aligned}$$

where we used results from the previous section (together with the sharpening) and the fact that $\mathbf{w}_j = \mathbf{z}_j\beta_j^{OLS} + \tilde{\boldsymbol{\epsilon}}$. For $\hat{\beta}_j^{HOLS}$, we analyse the numerator.

$$\sqrt{n}\frac{1}{n}\Big|\big(\mathbf{z}_j^3\big)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}} - \big(\hat{\mathbf{z}}_j^3\big)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}\Big| \leq \sqrt{n}\frac{1}{n}\Big|\big(\mathbf{z}_j^3 - \hat{\mathbf{z}}_j^3\big)^\top \tilde{\boldsymbol{\epsilon}}\Big| + \sqrt{n}\frac{1}{n}\Big|\big(\mathbf{z}_j^3 - \hat{\mathbf{z}}_j^3\big)^\top \boldsymbol{P}_{-j}\tilde{\boldsymbol{\epsilon}}\Big|. \tag{3.21}$$

Using a derivation as in (3.19) and (A3.3), we know $\|\boldsymbol{P}_{-j}\tilde{\boldsymbol{\epsilon}}\|_2 = \mathcal{O}_p(1)$. Thus, using (3.20), the second term is controlled. In (3.20), we have split $\Big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\Big)$ in to three parts. Only the third part is critical concerning the convergence of the first term in (3.21) as the others lead to $\mathcal{o}_p(1)$ terms when applying Cauchy-Schwarz to the inner product. Therefore, we take a closer look at $\Big(\mathbf{z}_j^2 \odot (\mathbf{z}_j - \hat{\mathbf{z}}_j)\Big)^\top \tilde{\boldsymbol{\epsilon}}$.

$$\begin{aligned}
&\sqrt{n}\frac{1}{n}\Big|\big(\mathbf{z}_j^2 \odot (\mathbf{z}_j - \hat{\mathbf{z}}_j)\big)^\top \tilde{\boldsymbol{\epsilon}}\Big| = \sqrt{n}\frac{1}{n}\left|\sum_{i=1}^n z_{ij}^2(z_{ij} - \hat{z}_{ij})\tilde{\epsilon}_i\right| = \sqrt{n}\frac{1}{n}\left|\sum_{i=1}^n z_{ij}^2 \mathbf{x}_{i,-j}\big(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\big)\tilde{\epsilon}_i\right| \\
=&\sqrt{n}\frac{1}{n}\left|\sum_{i=1}^n z_{ij}^2\tilde{\epsilon}_i \sum_{k\neq j} x_{ik}(\hat{\gamma}_{jk} - \gamma_{jk})\right| = \left|\sum_{k\neq j}(\hat{\gamma}_{jk} - \gamma_{jk})\sqrt{n}\frac{1}{n}\sum_{i=1}^n z_{ij}^2 x_{ik}\tilde{\epsilon}_i\right|
\end{aligned}$$

$$\leq \sum_{k\neq j} |\hat{\gamma}_{jk} - \gamma_{jk}| \left| \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 x_{ik}\tilde{\epsilon}_i \right| = \sum_{k\neq j} \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right)\mathcal{O}_p\big(\sqrt{n}\big)\mathcal{O}_p(1) = \mathcal{O}_p(1).$$

In the second to last inequality, we use (B3.1). In short,

$$\begin{aligned}
&\sqrt{n}\frac{1}{n}\left|\left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}} - \left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}\right| = \mathcal{O}_p(1) \text{ such that} \\
&\sqrt{n}\frac{1}{n}\left|\left(\tilde{\mathbf{z}}_j^3\right)^\top \tilde{\boldsymbol{\epsilon}} - \left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}\right| \\
\leq &\sqrt{n}\frac{1}{n}\left|\left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}} - \left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}\right| + \sqrt{n}\frac{1}{n}\left|\left(\tilde{\mathbf{z}}_j^3\right)^\top \tilde{\boldsymbol{\epsilon}} - \left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}\right| \\
= &\sqrt{n}\frac{1}{n}\left|\left(\mathbf{z}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}} - \left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}\right| + \sqrt{n}\frac{1}{n}\left|\left(\tilde{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j} \tilde{\boldsymbol{\epsilon}}\right| = \mathcal{O}_p(1).
\end{aligned}$$

This leads to

$$\begin{aligned}
\sqrt{n}\hat{\beta}_j^{HOLS} &= \frac{\sqrt{n}\frac{1}{n}\left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{w}}_j}{\frac{1}{n}\left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{z}}_j} = \sqrt{n}\beta_j^{OLS} + \frac{\sqrt{n}\frac{1}{n}\left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \tilde{\boldsymbol{\epsilon}}}{\frac{1}{n}\left(\hat{\mathbf{z}}_j^3\right)^\top \hat{\mathbf{z}}_j} \\
&= \sqrt{n}\beta_j^{OLS} + \frac{\sqrt{n}\frac{1}{n}\left(\tilde{\mathbf{z}}_j^3\right)^\top \tilde{\boldsymbol{\epsilon}} + \mathcal{O}_p(1)}{\mathbb{E}\Big[Z_j^4\Big] + \mathcal{O}_p(1)} = \sqrt{n}\beta_j^{OLS} + \frac{\sqrt{n}\frac{1}{n}\left(\tilde{\mathbf{z}}_j^3\right)^\top \tilde{\boldsymbol{\epsilon}}}{\mathbb{E}\Big[Z_j^4\Big]} + \mathcal{O}_p(1).
\end{aligned}$$

Combining the results for $\sqrt{n}\hat{\beta}_j^{OLS}$ and $\sqrt{n}\hat{\beta}_j^{HOLS}$, we find

$$\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big) = \sqrt{n}\frac{1}{n}\left(\frac{\left(\tilde{\mathbf{z}}_j^3\right)^\top}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j^\top}{\mathbb{E}\Big[Z_j^2\Big]}\right)\tilde{\boldsymbol{\epsilon}} + \mathcal{O}_p(1). \tag{3.22}$$

Since the first term is a scaled sum of i.i.d. random variables, we can apply the CLT to it

$$\sqrt{n}\frac{1}{n}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}} \xrightarrow{\mathbb{D}} \mathcal{N}\left(0, \text{Var}\left(\frac{\tilde{Z}_j^3\tilde{\mathcal{E}}}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{Z_j\tilde{\mathcal{E}}}{\mathbb{E}\Big[Z_j^2\Big]}\right)\right). \tag{3.23}$$

Note that $\mathbb{E}\Big[\tilde{Z}_j^3\tilde{\mathcal{E}}\Big] = 0$ as $\beta_j^{OLS} = \beta_j^{HOLS}$. Combining (3.22) and (3.23) leads to the first statement in Lemma 3.2. Applying the independence relationship induced by (B3.2) and (B3.3), the second statement follows trivially.

3.B.4 Proof of Theorem 3.5

From (3.22), we know $\sqrt{n}\frac{1}{n}\left|\hat{\mathbf{v}}_j^\top \tilde{\boldsymbol{\epsilon}} - \mathbf{v}_j^\top \tilde{\boldsymbol{\epsilon}}\right| \xrightarrow{\mathbb{P}} 0 \ \forall j \in U'$ under the given assumptions. For fixed dimensions, we can easily make this statement multivariate, i.e., $\sqrt{n}\frac{1}{n}\left\|\hat{\mathbf{v}}_U^\top \tilde{\boldsymbol{\epsilon}} - \mathbf{v}_U^\top \tilde{\boldsymbol{\epsilon}}\right\| \xrightarrow{\mathbb{P}} 0$. Therefore,

we inspect $\mathbf{v}_U^\top \tilde{\boldsymbol{\epsilon}}$ in the following. Note that this is a (scaled) sum of mean 0 i.i.d random vectors. Obviously, this enables the multivariate CLT such that

$$\sqrt{n}\frac{1}{n}\mathbf{v}_U^\top \tilde{\boldsymbol{\epsilon}} \overset{\mathbb{D}}{\to} \mathcal{N}\Big(\mathbf{0}, \mathbb{E}\Big[\tilde{\mathcal{E}}\mathbf{V}_U\mathbf{V}_U^\top\tilde{\mathcal{E}}\Big]\Big) = \mathcal{N}\Big(\mathbf{0}, \sigma_{\tilde{\mathcal{E}}}^2\mathbb{E}\Big[\mathbf{V}_U\mathbf{V}_U^\top\Big]\Big),$$

which implies the first part of Theorem 3.5. For the second part, note

$$\begin{aligned}
\frac{1}{n}\hat{\mathbf{v}}_j^\top\hat{\mathbf{v}}_k &= \frac{1}{n}\left(\frac{\boldsymbol{P}_{-j}^\perp\big(\hat{\mathbf{z}}_j^3\big)}{\frac{1}{n}\big(\hat{\mathbf{z}}_j^2\big)^\top\big(\hat{\mathbf{z}}_j^2\big)} - \frac{\hat{\mathbf{z}}_j}{\frac{1}{n}\hat{\mathbf{z}}_j^\top\hat{\mathbf{z}}_j}\right)^\top\left(\frac{\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big)}{\frac{1}{n}\big(\hat{\mathbf{z}}_k^2\big)^\top\big(\hat{\mathbf{z}}_k^2\big)} - \frac{\hat{\mathbf{z}}_k}{\frac{1}{n}\hat{\mathbf{z}}_k^\top\hat{\mathbf{z}}_k}\right)\\
&= \frac{1}{n}\left(\frac{\big(\hat{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big)}{\frac{1}{n}\big(\hat{\mathbf{z}}_j^2\big)^\top\big(\hat{\mathbf{z}}_j^2\big)\frac{1}{n}\big(\hat{\mathbf{z}}_k^2\big)^\top\big(\hat{\mathbf{z}}_k^2\big)} - \frac{\big(\hat{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\hat{\mathbf{z}}_k}{\frac{1}{n}\big(\hat{\mathbf{z}}_j^2\big)^\top\big(\hat{\mathbf{z}}_j^2\big)\frac{1}{n}\hat{\mathbf{z}}_k^\top\hat{\mathbf{z}}_k}\right.\\
&\quad\left. - \frac{\hat{\mathbf{z}}_j^\top\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big)}{\frac{1}{n}\hat{\mathbf{z}}_j^\top\hat{\mathbf{z}}_j\frac{1}{n}\big(\hat{\mathbf{z}}_k^2\big)^\top\big(\hat{\mathbf{z}}_k^2\big)} + \frac{\hat{\mathbf{z}}_j^\top\hat{\mathbf{z}}_k}{\frac{1}{n}\hat{\mathbf{z}}_j^\top\hat{\mathbf{z}}_j\frac{1}{n}\hat{\mathbf{z}}_k^\top\hat{\mathbf{z}}_k}\right)
\end{aligned}$$

For each of the denominator terms, convergence has been established already. For the numerator terms, we can apply (3.19), (3.20), and $\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2 = \mathcal{O}_p(1)$.

$$\begin{aligned}
&\frac{1}{n}\Big|\big(\hat{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big) - \big(\mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\mathbf{z}_k^3\big)\Big|\\
=&\frac{1}{n}\Big|\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big) + \big(\mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big) + \big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big)\Big|\\
\leq&\frac{1}{n}\Big|\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big)\Big| + \frac{1}{n}\Big|\big(\mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big)\Big| + \frac{1}{n}\Big|\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big)\Big|\\
\leq&\frac{1}{n}\Big\|\boldsymbol{P}_{-j}^\perp\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)\Big\|_2\Big\|\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big)\Big\|_2 + \frac{1}{n}\Big\|\boldsymbol{P}_{-j}^\perp\big(\hat{\mathbf{z}}_j^3\big)\Big\|_2\Big\|\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big)\Big\|_2 +\\
&\frac{1}{n}\Big\|\boldsymbol{P}_{-j}^\perp\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)\Big\|_2\Big\|\boldsymbol{P}_{-k}^\perp\big(\hat{\mathbf{z}}_k^3\big)\Big\|_2\\
\leq&\frac{1}{n}\big\|\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)\big\|_2\big\|\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big)\big\|_2 + \frac{1}{n}\big\|\hat{\mathbf{z}}_j^3\big\|_2\big\|\big(\hat{\mathbf{z}}_k^3 - \mathbf{z}_k^3\big)\big\|_2 + \frac{1}{n}\big\|\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)\big\|_2\big\|\hat{\mathbf{z}}_k^3\big\|_2\\
=&\mathcal{O}_p\left(\frac{K^2}{\sqrt{n}}\right) = o_p(1)\\
&\frac{1}{n}\Big|\big(\mathbf{z}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\mathbf{z}_k^3\big) - \big(\tilde{\mathbf{z}}_j^3\big)^\top\big(\tilde{\mathbf{z}}_k^3\big)\Big| = \frac{1}{n}\Big|\big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}^\perp\boldsymbol{P}_{-k}^\perp\big(\tilde{\mathbf{z}}_k^3\big) - \big(\tilde{\mathbf{z}}_j^3\big)^\top\big(\tilde{\mathbf{z}}_k^3\big)\Big|\\
=&\frac{1}{n}\Big|\big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}\boldsymbol{P}_{-k}\big(\tilde{\mathbf{z}}_k^3\big) + \big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-k}\big(\tilde{\mathbf{z}}_k^3\big) + \big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}\big(\tilde{\mathbf{z}}_k^3\big)\Big|\\
\leq&\frac{1}{n}\Big|\big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}\boldsymbol{P}_{-k}\big(\tilde{\mathbf{z}}_k^3\big)\Big| + \frac{1}{n}\Big|\big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-k}\big(\tilde{\mathbf{z}}_k^3\big)\Big| + \frac{1}{n}\Big|\big(\tilde{\mathbf{z}}_j^3\big)^\top\boldsymbol{P}_{-j}\big(\tilde{\mathbf{z}}_k^3\big)\Big|\\
\leq&\frac{1}{n}\big\|\boldsymbol{P}_{-j}\big(\tilde{\mathbf{z}}_j^3\big)\big\|_2\big\|\boldsymbol{P}_{-k}\big(\tilde{\mathbf{z}}_k^3\big)\big\|_2 + \frac{1}{n}\big\|\tilde{\mathbf{z}}_j^3\big\|_2\big\|\boldsymbol{P}_{-k}\big(\tilde{\mathbf{z}}_k^3\big)\big\|_2 + \frac{1}{n}\big\|\boldsymbol{P}_{-j}\big(\tilde{\mathbf{z}}_j^3\big)\big\|_2\big\|\tilde{\mathbf{z}}_k^3\big\|_2 = o_p(1) \text{ so}
\end{aligned}$$

$$\left(\hat{\mathbf{z}}_j^3\right)^\top \boldsymbol{P}_{-j}^\perp \boldsymbol{P}_{-k}^\perp \left(\hat{\mathbf{z}}_k^3\right) = \left(\tilde{\mathbf{z}}_j^3\right)^\top \left(\tilde{\mathbf{z}}_k^3\right) + \mathcal{O}_p(1) = \mathbb{E}\Big[\tilde{Z}_j^3 \tilde{Z}_k^3\Big] + \mathcal{O}_p(1)$$

The other terms follow in a very similar fashion such that Slutsky's theorem leads to

$$\frac{1}{n}\hat{\mathbf{v}}_j^\top \hat{\mathbf{v}}_k = \mathbb{E}[V_j V_k] + \mathcal{O}_p(1)$$

For fixed p, this can be directly made multidimensional which proves the theorem's statement.

3.B.4.1 Proof of Corollary 3.2

Consider $\mathbf{S}$ as given in Step 7 of Algorithm 3.1. Using the second part of Theorem 3.5 and a consistent estimate of $\hat{\sigma}$, we have

$$\sqrt{n}\mathbf{S}_{U'} \xrightarrow{\mathbb{D}} \mathcal{N}\Big(\mathbf{0}, \sigma_{\tilde{\mathcal{E}}}^2 \mathbb{E}\Big[\mathbf{V}_{U'}\mathbf{V}_{U'}^\top\Big]\Big).$$

Let $\mathbf{S}^* \sim \mathcal{N}\Big(\mathbf{0}, \sigma_{\tilde{\mathcal{E}}}^2 \mathbb{E}\big[\mathbf{V}_{U'}\mathbf{V}_{U'}^\top\big]\Big)$ and denote the cumulative density function (CDF) of its maximum absolute value by F^*. Denote the CDF of $\sqrt{n}\left\|\hat{\boldsymbol{\beta}}_{U'}^{HOLS} - \hat{\boldsymbol{\beta}}_{U'}^{OLS}\right\|_\infty$ by F_n. Let q be the quantile function and $\hat{q}$ the estimated quantile function using $\mathbf{s}^1, \ldots, \mathbf{s}^{n_{sim}}$. Then,

$$\begin{aligned}
&\lim_{n\to\infty} \lim_{n_{sim}\to\infty} \mathbb{P}\big(\exists j \in U' \text{ such that } H_{0,j} \text{ is rejected}\big) \\
=&\lim_{n\to\infty} \lim_{n_{sim}\to\infty} \mathbb{P}\Big(\left\|\hat{\beta}_{U'}^{HOLS} - \hat{\beta}_{U'}^{OLS}\right\|_\infty > \hat{q}_{1-\alpha}(\|\mathbf{S}\|_\infty)\Big) \\
\leq&\lim_{n\to\infty} \lim_{n_{sim}\to\infty} \mathbb{P}\Big(\left\|\hat{\beta}_{U'}^{HOLS} - \hat{\beta}_{U'}^{OLS}\right\|_\infty > \hat{q}_{1-\alpha}(\|\mathbf{S}_{U'}\|_\infty)\Big) = 1 - \lim_{n\to\infty} \lim_{n_{sim}\to\infty} F_n(\hat{q}_{1-\alpha}(\|\mathbf{S}_{U'}\|_\infty)) \\
=&1 - \lim_{n\to\infty} \lim_{n_{sim}\to\infty} F^*(\hat{q}_{1-\alpha}(\|\mathbf{S}_{U'}\|_\infty)) = 1 - \lim_{n\to\infty} F^*(q_{1-\alpha}(\|\mathbf{S}_{U'}\|_\infty)) = 1 - F^*(q_{1-\alpha}(\|\mathbf{S}^*\|_\infty)) \\
=&\alpha
\end{aligned}$$

For the equality between the second and third line, note that $F_n \to F$ using Theorem 3.5 and the continuous mapping theorem. As the maximum of several Gaussian random variables has a continuous CDF, this convergence is uniform such the convergence also holds at $\hat{q}_{1-\alpha}(\|\mathbf{S}_{U'}\|_\infty)$ which is not constant in n and n_{sim}. For the convergence of empirical quantiles, see, e.g., the discussion in (van der Vaart, 2000, Chapter 21).

3.B.5 Proof of Theorem 3.6

We split the goal into two problems, namely,

$$\lim_{n\to\infty} \mathbb{P}\Big[\hat{U} \subseteq U\Big] = 1 \text{ and } \lim_{n\to\infty} \mathbb{P}\Big[\hat{U} \supseteq U\Big] = 1. \tag{3.24}$$

The first one corresponds to rejecting the required $H_{0,j}$ and the second one to no wrong rejection. We provide the supporting lemmata. Theorem 3.6 follows directly by combining these.

Lemma 3.3. Assume that the data follows the model (3.11) and that (A3.1) - (A3.2) hold. Let $j \notin U$. Then,

$$\mathbb{P}[|t_j| \geq \tau_n] \geq \mathbb{P}\big[\big|\beta_j^{HOLS} - \beta_j^{OLS}\big| \geq \tau_n \big|\mathcal{O}_p\big(1/\sqrt{n}\big)\big| + |\mathcal{o}_p(1)|\big].$$

This probability can be ensured to approach 1 if we chose $\tau_n = \mathcal{o}(\sqrt{n})$. Under this condition, we find $\lim_{n\to\infty} \mathbb{P}\Big[\hat{U} \subseteq U\Big] = 1$. Notably, we assume $\Big|\beta_j^{HOLS} - \beta_j^{OLS}\Big|$ to be constant, i.e., we deal with a fixed alternative. We further remark that we could use a constant significance level $\alpha_n = \alpha$ to receive just the first convergence in (3.24).

Let us now turn to variables for which $H_{0,j}$ holds true. In order to reuse our convergence results from Section 3.3.1, we have to additionally invoke (A3.3), (B3.1) and (A3.4).

Lemma 3.4. Assume that the data follows the model (3.11) and that (A3.1) - (A3.3) hold. Let j be some covariate in U for which (B3.1) and (A3.4) hold. Then,

$$\mathbb{P}[|t_j| \geq \tau_n] \leq \mathbb{E}\left[\left(\left(\frac{\tilde{Z}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{Z_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)\tilde{\mathcal{E}}\right)^2\right] / \left(\tau_n |\mathcal{O}_p(1)|/2\right)^2 + \mathbb{P}[|\mathcal{o}_p(1)| \geq \tau_n/2].$$

Either term vanishes if we choose $1/\tau_n = \mathcal{o}(1)$. Thus, as long as τ_n grows at any rate, we receive $\lim_{n\to\infty} \mathbb{P}\Big[\hat{U} \supseteq U\Big] = 1$.

3.B.5.1 Proof of Lemma 3.3

From Theorem 3.3, we know

$$\begin{aligned}
\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big) &= \sqrt{n}\big(\beta_j^{HOLS} - \beta_j^{OLS}\big) + \mathcal{o}_p\big(\sqrt{n}\big) \\
\widehat{\mathrm{Var}}\Big(\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big)\Big) &= \mathcal{O}_p(1).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
|t_j| &= \left|\frac{\sqrt{n}\Big(\beta_j^{HOLS} - \beta_j^{OLS}\Big)}{\mathcal{O}_p(1)} + \mathcal{o}_p\big(\sqrt{n}\big)\right| \geq \left|\frac{\sqrt{n}\Big(\beta_j^{HOLS} - \beta_j^{OLS}\Big)}{\mathcal{O}_p(1)}\right| - \big|\mathcal{o}_p\big(\sqrt{n}\big)\big| \\
\mathbb{P}[|t_j| \geq \tau_n] &\geq \mathbb{P}\left[\left|\frac{\sqrt{n}\Big(\beta_j^{HOLS} - \beta_j^{OLS}\Big)}{\mathcal{O}_p(1)}\right| \geq \tau_n + \big|\mathcal{o}_p\big(\sqrt{n}\big)\big|\right] \\
&= \mathbb{P}\big[\big|\beta_j^{HOLS} - \beta_j^{OLS}\big| \geq \tau_n \big|\mathcal{O}_p\big(1/\sqrt{n}\big)\big| + |\mathcal{o}_p(1)|\big].
\end{aligned}$$

3.B.5.2 Proof of Lemma 3.4

For variables fulfilling (B3.1), we know from (3.22) and Theorem 3.3

$$\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big) = \sqrt{n}\frac{1}{n}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}} + \mathcal{O}_p(1).$$
$$\widehat{\mathrm{Var}}\Big(\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big)\Big) = \mathcal{O}_p(1).$$

This yields

$$\begin{aligned}
|t_j| &= \left|\sqrt{n}\frac{1}{n\mathcal{O}_p(1)}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}} + \mathcal{o}_p(1)\right| \\
&\leq \left|\sqrt{n}\frac{1}{n\mathcal{O}_p(1)}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}}\right| + |\mathcal{o}_p(1)| \\
\mathbb{P}[|t_j| \geq \tau_n] &\leq \mathbb{P}\left[\left|\sqrt{n}\frac{1}{n\mathcal{O}_p(1)}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}}\right| + |\mathcal{o}_p(1)| \geq \tau_n\right] \\
&\leq \mathbb{P}\left[\left|\sqrt{n}\frac{1}{n\mathcal{O}_p(1)}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}}\right| \geq \tau_n/2\right] + \mathbb{P}[|\mathcal{o}_p(1)| \geq \tau_n/2] \\
&\leq \mathbb{P}\left[\left|\sqrt{n}\frac{1}{n}\left(\frac{\tilde{\mathbf{z}}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{\mathbf{z}_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^\top \tilde{\boldsymbol{\epsilon}}\right| \geq \tau_n|\mathcal{O}_p(1)|/2\right] + \mathbb{P}[|\mathcal{o}_p(1)| \geq \tau_n/2] \\
&\leq \mathbb{E}\left[\left(\left(\frac{\tilde{Z}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \frac{Z_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)\mathcal{E}\right)^2\right]/\left(\tau_n|\mathcal{O}_p(1)|/2\right)^2 + \mathbb{P}[|\mathcal{o}_p(1)| \geq \tau_n/2],
\end{aligned}$$

where the last step follows from Chebyshev's inequality, assuming the second moment exists (cf. (A3.4)).

3.B.6 Proof of Theorem 3.8

From the definitions in (3.17), we see that $\beta_j = \beta_j^*$ iff $\left(\left(\boldsymbol{\omega}_{N,M}\boldsymbol{\omega}_{M,M}^{-1}\right)^\top \boldsymbol{\beta}_N^*\right)_j = 0$. We can inspect this further

$$\left(\left(\boldsymbol{\omega}_{N,M}\boldsymbol{\omega}_{M,M}^{-1}\right)^\top \boldsymbol{\beta}_N^*\right)_j = \left(\left(\boldsymbol{\omega}_{M,M}^{-1}\right)^\top \boldsymbol{\omega}_{N,M}^\top \boldsymbol{\beta}_N^*\right)_j = \left(\boldsymbol{\omega}_{M,M}^{-1}\right)_j^\top \sum_{k\in N} \boldsymbol{\omega}_{k,M}^\top \beta_k^* = \sum_{k\in N}\left(\boldsymbol{\omega}_{M,M}^{-1}\right)_j^\top \boldsymbol{\omega}_{k,M}^\top \beta_k^*.$$

For some variable $k \in N$, we have

$$\begin{aligned}
\mathbb{E}\Big[\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M(\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M)^\top\Big] &= \boldsymbol{\omega}_{M,M}\boldsymbol{\Sigma}^{\boldsymbol{\Psi}_M}\boldsymbol{\omega}_{M,M}^\top \quad \text{and} \\
\mathbb{E}\Big[\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M(\boldsymbol{\omega}_{k,M}\boldsymbol{\Psi}_M)^\top\Big] &= \boldsymbol{\omega}_{M,M}\boldsymbol{\Sigma}^{\boldsymbol{\Psi}_M}\boldsymbol{\omega}_{k,M}^\top. \quad \text{Thus,} \\
\mathbb{E}\Big[\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M(\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M)^\top\Big]^{-1}\mathbb{E}\Big[\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M(\boldsymbol{\omega}_{k,M}\boldsymbol{\Psi}_M)^\top\Big] &= \Big(\boldsymbol{\omega}_{M,M}^{-1}\Big)^\top\boldsymbol{\omega}_{k,M}^\top
\end{aligned}$$

is the regression parameter of the regression $\boldsymbol{\omega}_{k,M}\boldsymbol{\Psi}_M$ versus $\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M$. Naturally, $\boldsymbol{\omega}_{k,M}\boldsymbol{\Psi}_M$ can be perfectly recovered by a linear combination of $\boldsymbol{\omega}_{M,M}\boldsymbol{\Psi}_M$ using only k's nearest ancestors in M, say, $\mathrm{PA}^M(k)$. Thus, $\Big(\boldsymbol{\omega}_{M,M}^{-1}\Big)_j^\top\boldsymbol{\omega}_{k,M} = 0$ if $j \notin \mathrm{PA}^M(k)$. Extending this argument to all $k \in N$ the theorem's statement follows.

3.B.7 Proof of Theorem 3.9

We provide some supporting lemmata.

Lemma 3.5. Assume that the data follows the model (3.16) without hidden variables. Then,

$$Z_j = \delta_{j,j}\Psi_j + \sum_{k\in\mathrm{CH}(j)}\delta_{j,k}\Psi_k \quad j = 1,\dots,p$$

for an appropriate set of parameters. Further, the support of $\boldsymbol{\gamma}_j$ (cf. (3.5)) is restricted to j's Markov boundary.

Thus, only the "noise" of j itself or its children remains in Z_j.

Now, consider the best regression of Z_j^3 versus $\mathbf{X}_{-j}$ as defined in (3.7). We get an analogous result for the residuum $\tilde{Z}_j^3$.

Lemma 3.6. Assume that the data follows the model (3.16) without hidden variables. Then,

$$\tilde{Z}_j^3 = Z_j^3 + \tilde{\delta}_{j,j}\Psi_j + \sum_{k\in\mathrm{CH}(j)}\tilde{\delta}_{j,k}\Psi_k \quad j = 1,\dots,p$$

for an appropriate set of parameters. Further, the support of $\tilde{\boldsymbol{\gamma}}_j$ (cf. (3.7)) is restricted to j's Markov boundary.

Finally, we inspect the regression of Ψ_k versus $\mathbf{X}_M$ for some $k \in N$. With a slight abuse of notation, define

$$Z_k \coloneqq \Psi_k - \mathbf{X}_M^\top\boldsymbol{\beta}^k, \quad \text{where} \quad \boldsymbol{\beta}^k \coloneqq \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^{|M|}}\mathbb{E}\left[\Big(\Psi_k - \mathbf{X}_M^\top\mathbf{b}\Big)^2\right] = \mathbb{E}\Big[\mathbf{X}_M\mathbf{X}_M^\top\Big]^{-1}\mathbb{E}[\mathbf{X}_M\Psi_k]. \quad (3.25)$$

Lemma 3.7. Assume that the data follows the model (3.16). Let $\mathbf{X}_M$ and $\mathbf{X}_N$ be the observed and the hidden variables, where $k \in N$. Then,

$$Z_k = \sum_{l \in N} \delta_{k,l} \Psi_l + \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \Psi_m, \quad \text{where} \quad \mathrm{CH}_N = \left(\bigcup_{l \in N} \mathrm{CH}(l) \right) \setminus N,$$

for an appropriate set of parameters. Further, the support of $\boldsymbol{\beta}^k$ is restricted to the union of the hidden variables' Markov boundaries.

As $\tilde{\mathcal{E}}$ is a linear combination of these Z_k and the independent $\mathcal{E}$, we can combine Lemmata 3.5 and 3.6 and 3.7 to find $Z_j \perp \tilde{\mathcal{E}}$ and $\tilde{Z}_j^3 \perp \tilde{\mathcal{E}}$ for some variable j outside the hidden variables' Markov boundaries. Furthermore, $\boldsymbol{\beta}^{OLS} - \boldsymbol{\beta}$ is a linear combination of the $\boldsymbol{\beta}^k$ for $k \in N$ such that outside the hidden variables' Markov boundaries the two parameters are equal as claimed.

To check (B3.1), split X_k into a part consisting of Ψ_j and Ψ_l $\forall l \in \mathrm{CH}(j)$, say $X_{k,1}$, independent from $\tilde{\mathcal{E}}$, and the remainder, say $X_{k,2}$, independent from Z_j. Then, we find

$$\begin{aligned} \mathbb{E}\Big[Z_j^2 X_k \tilde{\mathcal{E}}\Big] &= \mathbb{E}\Big[Z_j^2 X_{k,1} \tilde{\mathcal{E}}\Big] + \mathbb{E}\Big[Z_j^2 X_{k,2} \tilde{\mathcal{E}}\Big] = \mathbb{E}\big[Z_j^2 X_{k,1}\big] \mathbb{E}\Big[\tilde{\mathcal{E}}\Big] + \mathbb{E}\big[Z_j^2\big] \mathbb{E}\Big[X_{k,2} \tilde{\mathcal{E}}\Big] \\ &= \mathbb{E}\big[Z_j^2\big] \mathbb{E}\Big[X_{k,2} \tilde{\mathcal{E}}\Big] = \mathbb{E}\big[Z_j^2\big] \Big(\mathbb{E}\Big[X_k \tilde{\mathcal{E}}\Big] - \mathbb{E}\Big[X_{k,1} \tilde{\mathcal{E}}\Big] \Big) = 0. \end{aligned}$$

3.B.7.1 Proof of Lemma 3.5

Recall the representation

$$Z_j = \delta_{j,j} \Psi_j + \sum_{k \in \mathrm{CH}(j)} \delta_{j,k} \Psi_k. \tag{3.26}$$

Assume first only the noise terms Ψ_j and Ψ_k $\forall k \in \mathrm{CH}_j$ exist, while all the other terms are set to 0. Call the variables in this construction X'_k and the residuum Z'_j. Obviously, Z'_j has a representation as in (3.26). Now by the definition of least squares, Z'_j and Z_j always have the smallest possible variance in their given model. If we add more independent noise terms to the model, the variance cannot decrease. Therefore, it holds $\mathrm{Var}(Z_j) \geq \mathrm{Var}\Big(Z'_j\Big)$. Thus, if there exists a parameter such that $X_j - \boldsymbol{\gamma}^\top \mathbf{X}_{-k} = Z'_j$, it must be optimal such that $Z_j = Z'_j$. Let now $\boldsymbol{\gamma} = -\boldsymbol{\delta}_j$. Then, we have

$$\begin{aligned} X_j - \sum_{k \neq j} \gamma_k X_k = X_j + \sum_{k \in \mathrm{CH}_j} \delta_{j,k} X_k &= X_j + \sum_{k \in \mathrm{CH}(j)} \delta_{j,k} \left(\Psi_k + \theta_{k,j} X_j + \sum_{l \in \mathrm{PA}(k) \setminus j} \theta_{k,l} X_l \right) \\ &= X_j \left(1 + \sum_{k \in \mathrm{CH}(j)} \delta_{j,k} \theta_{k,j} \right) + \sum_{k \in \mathrm{CH}(j)} \delta_{j,k} \Psi_k + \sum_{k \in \mathrm{CH}(j)} \delta_{j,k} \sum_{l \in \mathrm{PA}(k) \setminus j} \theta_{k,l} X_l. \end{aligned}$$

Now adjust $\boldsymbol{\gamma}$ by

- $\forall l \in \mathrm{PA}(j)$ adding $\Big(1 + \sum_{k\in\mathrm{CH}(j)} \delta_{j,k}\theta_{k,j}\Big)\theta_{j,l}$ to γ_l

- $\forall k \in \mathrm{CH}(j), \quad \forall l \in \mathrm{PA}(k) \setminus j$ adding $\delta_{j,k}\theta_{k,l}$ to γ_l

This leads to

$$X_j - \sum_{k\neq j} \gamma_k X_k = \Psi_j \left(1 + \sum_{k\in\mathrm{CH}(j)} \delta_{j,k}\theta_{k,j}\right) + \sum_{k\in\mathrm{CH}(j)} \delta_{j,k}\Psi_k.$$

This is almost the optimal Z'_j as in (3.26). It remains to argue that the term in the bracket equals $\delta_{j,j}$. For this, note that the weighted sum of terms that include Ψ_k in the construction of Z'_j must be exactly $\delta_{j,k}$. These terms can occur from adding a multiple of either k itself or of descendants thereof (that are children of j as well). These descendants have "inherited" the same multiple of Ψ_k as of $\theta_{k,j}\Psi_j$. Therefore, there is a net contribution of $\delta_{j,k}\theta_{k,j}\Psi_j$ originating from variable k. Applying this argument to each child, we receive the desired sum. Naturally, the 1 is the contribution from $X'_j = \Psi_j$ itself.

Thus, we receive the desired construction of Z_j. Further, we see that in this construction the support of $\boldsymbol{\gamma}_j$ is restricted to j's parents, its children, and its children's other parents, which is exactly the second part of the lemma.

3.B.7.2 Proof of Lemma 3.6

This follows using very similar arguments as in Section 3.B.7.1 and is omitted here for simplicity.

3.B.7.3 Proof of Lemma 3.7

Recall the construction

$$Z_k = \sum_{l\in N} \delta_{k,l}\Psi_l + \sum_{m\in\mathrm{CH}_N} \delta_{k,m}\Psi_m, \quad \text{where} \quad \mathrm{CH}_N = \left(\bigcup_{l\in N} \mathrm{CH}(l)\right) \setminus N. \tag{3.27}$$

We argue as before: Assume first only these variables are nonzero leading to an optimal residuum Z'_k which has a representation as in (3.27). Naturally, $Z_k = \Psi_k - \left(\boldsymbol{\beta}^k\right)^\top \mathbf{X}_M$ for some $\boldsymbol{\beta}^k$. If we find a parameter such that $\Psi_k - \left(\boldsymbol{\beta}^k\right)^\top \mathbf{X}_M = Z'_k$, it must be optimal as $\mathrm{Var}(Z_k) \geq \mathrm{Var}(Z'_k)$. We now construct such a parameter. Start by $\boldsymbol{\beta}^k = -\boldsymbol{\delta}_k$.

$$\begin{aligned}
-\left(\boldsymbol{\beta}^k\right)^\top \mathbf{X}_M &= \sum_{m\in\mathrm{CH}_N} \delta_{k,m} X_m = \sum_{m\in\mathrm{CH}_N} \delta_{k,m}(\mathcal{E}_{X_m} + \boldsymbol{\omega}_{m,N}\boldsymbol{\Psi}_N) \\
&= \sum_{l\in N} \Psi_l \sum_{m\in\mathrm{CH}_N} \delta_{k,m}\omega_{m,l} + \sum_{m\in\mathrm{CH}_N} \delta_{k,m}\mathcal{E}_{X_m} \\
&= \sum_{l\in N} \Psi_l \sum_{m\in\mathrm{CH}_N} \delta_{k,m}\omega_{m,l} + \sum_{m\in\mathrm{CH}_N} \delta_{k,m}\left(\Psi_m + \sum_{r\in\mathrm{PA}^M(m)} \theta'_{mr}\mathcal{E}_{X_r}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l \in N} \Psi_l \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \omega_{m,l} + \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \left(\Psi_m + \sum_{r \in \mathrm{PA}^M(m)} \theta'_{mr} (X_r - \boldsymbol{\omega}_{r,H} \boldsymbol{\Psi}_H) \right) \\
&= \sum_{l \in N} \Psi_l \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \left(\omega_{m,l} - \sum_{r \in \mathrm{PA}^M(m)} \theta'_{mr} \omega_{r,l} \right) \\
&+ \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \left(\Psi_m + \sum_{r \in \mathrm{PA}^M(m)} \theta'_{mr} X_r \right),
\end{aligned}
$$

where we have used the fact that $\mathcal{E}_{\boldsymbol{X}}$ follows a linear SEM as well for some suitable set of parameters $\theta'_{m,r}$. Now adjust $\boldsymbol{\beta}^k$ by

- $\forall m \in \mathrm{CH}_N, \quad \forall r \in \mathrm{PA}^M(m)$ adding $\delta_{k,m} \theta'_{m,r}$ to β^k_r.

This leads to

$$
\Psi_k - \left(\boldsymbol{\beta}^k \right)^\top \mathbf{X}_M = \Psi_k + \sum_{l \in N} \Psi_l \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \left(\omega_{m,l} - \sum_{r \in \mathrm{PA}^M(m)} \theta'_{mr} \omega_{r,l} \right) + \sum_{m \in \mathrm{CH}_N} \delta_{k,m} \Psi_m,
$$

which is as in (3.27). It remains to argue that the coefficient for Ψ_l equals $\delta_{k,l}$. Note that $\forall m \in \mathrm{CH}_N$ there is a net contribution of $\delta_{k,m} \Psi_m$ coming from a weighted sum of m and its descendants. There must be an according net contribution of all other parts that X'_m does not inherit from its parents (in M). These must be multiples of Ψ_l for $l \in N$. If $\omega_{r,l} \neq 0$, there is already a multiple of Ψ_l in X'_r. Thus, X'_m inherits $\theta'_{mr} \omega_{r,l}$ from X'_r. Extending this argument to all parents, a total of $\sum_{r \in \mathrm{PA}^M(m)} \theta'_{mr} \omega_{r,l}$ is inherited. The remainder, i.e., $\omega_{m,l} - \sum_{r \in \mathrm{PA}^M(m)} \theta'_{mr} \omega_{r,l}$ must then originate from X'_k itself and there is a contribution of $\delta_{k,m}$ times this remainder times Ψ_l. As this holds $\forall m \in \mathrm{CH}_N$, one can add all contributions leading to the desired sum.

Thus, we have established that $Z_k = Z'_k$ such that Z_k and $\boldsymbol{\beta}^k$ must be optimal. We also see that the support of $\boldsymbol{\beta}^k$ is restricted to CH_N and $\bigcup_{m \in \mathrm{CH}_N} \mathrm{PA}^M(m)$, which is exactly the hidden variables' Markov boundary.

3.C Theory for the high-dimensional extension

We provide additional details on the high-dimensional extension including proofs of the main results.

To understand Theorem 3.2, take a closer look at the difference between $\hat{\beta}^{OLS}_j$ and $\hat{\beta}^{HOLS}_j$, which

can be written as

$$\begin{aligned}\sqrt{n}\Big(\hat{\beta}_j^{HOLS} - \hat{\beta}_j^{OLS}\Big) =& \sqrt{n}\frac{\Big(\hat{\hat{\mathbf{z}}}_j^3\Big)^\top \mathbf{x}_{-j}/n}{\Big(\hat{\hat{\mathbf{z}}}_j^3\Big)^\top \mathbf{x}_j/n}\Big(\boldsymbol{\beta}_{-j} - \hat{\boldsymbol{\beta}}_{-j}\Big) - \sqrt{n}\frac{\hat{\mathbf{z}}_j^\top \mathbf{x}_{-j}/n}{\hat{\mathbf{z}}_j^\top \mathbf{x}_j/n}\Big(\boldsymbol{\beta}_{-j} - \hat{\boldsymbol{\beta}}_{-j}\Big) + \\ & \left(\frac{\Big(\hat{\hat{\mathbf{z}}}_j^3\Big)^\top/\sqrt{n}}{\Big(\hat{\hat{\mathbf{z}}}_j^3\Big)^\top \mathbf{x}_j/n} - \frac{\hat{\mathbf{z}}_j^\top/\sqrt{n}}{\hat{\mathbf{z}}_j^\top \mathbf{x}_j/n}\right)\boldsymbol{\epsilon} \coloneqq \Delta_j^{HOLS} - \Delta_j^{OLS} + \sqrt{n}\frac{1}{n}\hat{\mathbf{v}}_j^\top \boldsymbol{\epsilon}\end{aligned} \tag{3.28}$$

$$\sqrt{n}\Big(\hat{\boldsymbol{\beta}}^{HOLS} - \hat{\boldsymbol{\beta}}^{HOLS}\Big) = \boldsymbol{\Delta}^{HOLS} - \boldsymbol{\Delta}^{OLS} + \sqrt{n}\frac{1}{n}\hat{\mathbf{v}}^\top \boldsymbol{\epsilon}. \tag{3.29}$$

Thus, it consists of two bias terms and an error term, whose variance we can estimate. In the following, we inspect the terms in (3.28) assuming model (3.4) to justify Algorithm 3.2 and Theorem 3.2. Δ_j^{OLS} is under control under certain conditions as discussed in (van de Geer et al., 2014). For Δ_j^{HOLS} as well as the error scaling, we invoke our extra assumptions.

We consider the bias term.

Lemma 3.8. Assume that the data follows the model (3.4) with sub-Gaussian $\mathcal{E}$ and that (C3.1) - (C3.3) and (C3.5) - (C3.6) hold $(\forall j)$. Let $\hat{\boldsymbol{\beta}}$ come from Lasso regression with $\lambda \asymp \sqrt{\log(p)/n}$, $\hat{\mathbf{z}}_j$ from nodewise Lasso regression using $\lambda_j \asymp \sqrt{\log(p)/n}$, and $\hat{\hat{\mathbf{z}}}_j^3$ from nodewise Lasso regression of $\hat{\mathbf{z}}_j^3$ versus $\mathbf{x}_{-j}$ using $\tilde{\lambda}_j \asymp \max\Big\{\log(p)^{5/2}n^{-1/2}, s_j^2\log(p)^{5/2}n^{-3/2}, s_j\log(p)^2 n^{-1}, \sqrt{s_j}\log(p)n^{-1/2}\Big\}$. Use the definitions in (3.28). Then,

$$\big\|\boldsymbol{\Delta}^{HOLS}\big\|_\infty = \mathcal{O}_p(1) \quad \text{and} \quad \big\|\boldsymbol{\Delta}^{OLS}\big\|_\infty = \mathcal{O}_p(1).$$

Thus, under suitable assumptions, the bias vanishes. To get powerful tests, we want the variance of the error term to stay bounded. For asymptotically valid tests, we must ensure that the estimated standard deviation is of higher order of magnitude than the bias term.

Lemma 3.9. Assume that the data follows the model (3.4) with sub-Gaussian $\mathcal{E}$ and that (C3.1), (C3.2), (C3.4) and (C3.6) hold $(\forall j)$. Let $\hat{\boldsymbol{\beta}}$ come from Lasso regression with $\lambda \asymp \sqrt{\log(p)/n}$, $\hat{\mathbf{z}}_j$ from nodewise Lasso regression using $\lambda_j \asymp \sqrt{\log(p)/n}$, and $\hat{\hat{\mathbf{z}}}_j^3$ from nodewise Lasso regression of $\hat{\mathbf{z}}_j^3$ versus $\mathbf{x}_{-j}$ using $\tilde{\lambda}_j \asymp \max\Big\{\log(p)^{5/2}n^{-1/2}, s_j^2\log(p)^{5/2}n^{-3/2}, s_j\log(p)^2 n^{-1}, \sqrt{s_j}\log(p)n^{-1/2}\Big\}$. Use the definitions in (3.28). Then,

$$\frac{1}{n}\hat{\mathbf{v}}_j^\top \hat{\mathbf{v}}_j \xrightarrow{\mathbb{P}} \mathbb{E}\big[V_j^2\big] \quad \text{uniformly in } j$$

where $V_j = \dfrac{\tilde{Z}_j^3}{\mathbb{E}\Big[Z_j^4\Big]} - \dfrac{Z_j}{\mathbb{E}\Big[Z_j^2\Big]}$.

Note that the sub-Gaussian assumption ensures $\mathbb{E}\Big[V_j^2\Big] < \infty$, while as $\mathbb{E}\Big[V_j^2\Big] > 0$ if $\mathbb{E}\Big[\Big(\tilde{Z}_j^3\Big)^2\Big]\mathbb{E}\Big[Z_j^2\Big] > \mathbb{E}\Big[Z_j^4\Big]^2$, which always holds if $\tilde{Z}_j^3$ is not a linear function of Z_j. Thus, the estimate of the standard error approaches a bounded positive constant, enabling asymptotically valid z-tests.

In (C3.3), we have joint conditions on the different sparsity levels s and s_j. Thus, the larger the one is, the more restrictive the assumption on the other is. Let us consider some specific cases, namely, $s \approx s_j$, s_j maximal according to (C3.4), and s maximal according to (C3.2).

$$\text{For } s \approx s_j, \text{ we need } s \approx s_j = \mathcal{O}\left(\frac{n^{1/3}}{\log(p)}\right).$$
$$\text{For } s_j = \mathcal{O}\left(\frac{n^{3/5}}{\log(p)}\right), \text{ we need } s = \mathcal{O}\left(\frac{n^{1/5}}{\log(p)}\right).$$
$$\text{For } s = \mathcal{O}\left(\frac{n^{1/2}}{\log(p)^3}\right), \text{ we need } s_j = \mathcal{O}\Big(\log(p)^3\Big).$$

Note that if $\tilde{\lambda}_j$ is chosen optimally with respect to s_j, (C3.5) is actually the same as (C3.2) and (C3.3) such that there is no extra assumption on s that one has to invoke. In (C3.6), we have joint conditions on the different sparsity levels s_j and $\tilde{s}_j$. Thus, the larger the one is, the more restrictive the assumption on the other is. Let us consider some specific cases, namely, $s_j \approx \tilde{s}_j$, s_j maximal according to (C3.4), and $\tilde{s}_j$ maximal according to (C3.6).

$$\text{For } s_j \approx \tilde{s}_j, \text{ we need } s_j \approx \tilde{s}_j = \mathcal{O}\left(\frac{n^{1/2}}{\log(p)}\right).$$
$$\text{For } s_j = \mathcal{O}\left(\frac{n^{3/5}}{\log(p)}\right), \text{ we need } \tilde{s}_j = \mathcal{O}\left(\frac{n^{2/5}}{\log(p)}\right).$$
$$\text{For } \tilde{s}_j = \mathcal{O}\left(\frac{n}{\log(p)^5}\right), \text{ we need } s_j = \mathcal{O}\Big(\log(p)^3\Big).$$

Naturally, s_j and $\tilde{s}_j$ are to some extent related. In a linear SEM, the support of $\boldsymbol{\gamma}_j$ and the support of $\tilde{\boldsymbol{\gamma}}_j$ always lie within j's Markov boundary as we argue in Section 3.4.2. For completely arbitrary setups, it is typically even all of the boundary. Thus, $s_j = \tilde{s}_j$ would then be usual, except for "sink" nodes, such that the first case appears to be most interesting. Furthermore, if $\mathbb{E}\Big[Z_j^3 \mathbf{X}_{-j}\Big] = \mathbf{0}\ \forall j$, it holds $\tilde{s}_j = 0$ and (C3.6) is automatically fulfilled.

3.C.1 Proof of Lemma 3.8

Following the proofs in van de Geer et al. (2014), the assumptions are sufficient to claim

$$\|\mathbf{\Delta}^{OLS}\|_\infty = \mathcal{O}_p(1), \quad \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1 = \mathcal{O}_p(s\lambda), \quad \frac{1}{n}\left\|\mathbf{x}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\|_2^2 = \mathcal{O}_p\big(s\lambda^2\big),$$
$$\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_1 = \mathcal{O}_p(s_j\lambda_j)\ \forall j \quad \text{and} \quad \frac{1}{n}\|\mathbf{x}_{-j}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)\|_2^2 = \mathcal{O}_p\big(s_j\lambda_j^2\big)\ \forall j.$$

We now turn to $\|\mathbf{\Delta}^{HOLS}\|_\infty$. To control this, we want to ensure that

$$\left|\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_{-j}\left(\boldsymbol{\beta}_{-j} - \hat{\boldsymbol{\beta}}_{-j}\right)\right|/n = \mathcal{O}_p\big(1/\sqrt{n}\big) \quad \text{and} \quad \left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_j/n = \mathbb{E}\big[Z_j^4\big] + \mathcal{O}_p(1). \tag{3.30}$$

Note that we always have $\left\|\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_{-j}\right\|_\infty/n = \tilde{\lambda}_j$. Thus, the first goal in (3.30) is fulfilled using (C3.2) and (C3.3). Further,

$$\hat{\mathbf{z}}_j^3 = \mathbf{z}_j^3 + \big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big) = \mathbf{x}_{-j}\tilde{\boldsymbol{\gamma}}_j + \tilde{\mathbf{z}}_j^3 + \big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big).$$

From standard Lasso theory (cf. Bühlmann and van de Geer (2011)), we know that the order of which we should choose the tuning parameter $\tilde{\lambda}_j$ is dependent on the bound for

$$\begin{aligned}\frac{1}{n}\big\|\mathbf{x}_{-j}\big(\tilde{\mathbf{z}}_j^3 + \big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)\big)\big\|_\infty \leq& \frac{1}{n}\big\|\mathbf{x}_{-j}\tilde{\mathbf{z}}_j^3\big\|_\infty + \frac{1}{n}\big\|\mathbf{x}_{-j}\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)\big\|_\infty \\ \leq& \frac{1}{n}\big\|\mathbf{x}_{-j}\tilde{\mathbf{z}}_j^3\big\|_\infty + \frac{1}{n}\|\mathbf{x}_{-j}\|_\infty\big\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big\|_1.\end{aligned}$$

For the first term,

$$\begin{aligned}\mathbb{E}\left[\max_j\left\|\big(\tilde{\mathbf{z}}_j^3\big)^\top \mathbf{x}_{-j}\right\|_\infty\right] =& \mathbb{E}\left[\max_{j,k\neq j}\left|\big(\tilde{\mathbf{z}}_j^3\big)^\top \mathbf{x}_k\right|\right] = \mathbb{E}\left[\max_{j,k\neq j}\left|\sum_{i=1}^n \tilde{z}_{ij}^3 x_{ik}\right|\right] \\ =& \mathbb{E}\left[\max_{j,k\neq j}\left|\sum_{i=1}^n \tilde{z}_{ij}^3 x_{ik} - \mathbb{E}\big[\tilde{z}_{ij}^3 x_{ik}\big]\right|\right].\end{aligned}$$

We maximize over j as well to receive results uniformly in j. In the last equality, we use $\mathbb{E}\Big[\tilde{z}_{ij}^3 x_{ik}\Big] = \mathbb{E}\Big[\tilde{Z}_j^3 X_k\Big] = 0$.

The terms of the type $\tilde{z}_{ij}^3 x_{ik}$ can be viewed as different functions of the vector $\begin{pmatrix} x_{i1} & \dots & x_{ip} & \tilde{z}_{i1}^3 & \dots & \tilde{z}_{ip}^3 \end{pmatrix}^\top$. In total, these are $p(p-1)$ functions. For these, we can apply the Nemirovski moment inequality from Lemma 14.24 in Bühlmann and van de Geer (2011), which yields

$$\mathbb{E}\max_{j,k}\left|\sum_{i=1}^n \tilde{z}_{ij}^3 x_{ik} - \mathbb{E}\big[\tilde{z}_{ij}^3 x_{ik}\big]\right| \leq \big(8\log(2p(p-1))\big)^{1/2}\ \mathbb{E}\left[\max_{j,k}\sum_{i=1}^n\big(\tilde{z}_{ij}^3 x_{ik}\big)^2\right]^{1/2} \leq$$

$$(8\log(2p(p-1)))^{1/2}\ \mathbb{E}\left[\sum_{i=1}^{n}\max_{j,k}\left(\tilde{z}_{ij}^3\right)^2 x_{ik}^2\right]^{1/2} = (8\log(2p(p-1))n)^{1/2}\ \mathbb{E}\left[\max_{j,k}\left(\tilde{Z}_j^3\right)^2 X_k^2\right]^{1/2}$$
$$\leq (8\log(2p(p-1))n)^{1/2}\ \mathbb{E}\left[\max_k X_k^8\right]^{1/2}.$$

In the last expression, we simplify the notation and let $k \in \{1,\dots,2p\}$ with $X_{p+j} = \left(\tilde{Z}_j^3\right)^{1/3}$. We aim to bound that last expectation term, for which we use the sub-Gaussian assumption.

$$\begin{aligned}
\mathbb{E}\left[\max_k X_k^8\right] &= \int_0^\infty \mathbb{P}\left(\max_k X_k^8 > t\right)dt = \int_0^\infty \mathbb{P}\left(\max_k |X_k| > t^{1/8}\right)dt \\
&\leq \int_0^\infty \min\left\{1, \sum_k \mathbb{P}\left(|X_k| > t^{1/8}\right)\right\}dt \leq \int_0^\infty \min\left\{1, 2p\ \max_k \mathbb{P}\left(|X_k| > t^{1/8}\right)\right\}dt \\
&\leq \int_0^\infty \min\left\{1, 4p\ \max_k \exp\left(-\frac{t^{1/4}}{2\sigma_k^2}\right)\right\}dt \leq \int_0^a 1dt + 4p\int_a^\infty \exp\left(-\frac{t^{1/4}}{2\sigma_{\max}^2}\right)dt \\
&= a + p\ \exp\left(-\frac{a^{1/4}}{2\sigma_{\max}^2}\right)\text{poly}(a) = a + \exp\left(-\frac{a^{1/4}}{2\sigma_{\max}^2} + \log(p)\right)\text{poly}(a)
\end{aligned}$$

This holds for any positive integration bound a. If we choose $a > 16\sigma_{\max}^8\log(p)^4$, the second term will vanish as $p \to \infty$ leading to

$$\begin{aligned}
\mathbb{E}\left[\max_k X_k^8\right] &\leq \mathcal{O}\left(\log(p)^4\right) \text{ such that} \\
\mathbb{E}\left[\max_j \frac{1}{n}\left\|\left(\tilde{\mathbf{z}}_j^3\right)^\top \mathbf{x}_{-j}\right\|_\infty\right] &\leq \frac{1}{n}(8\log(2p(p-1))n)^{1/2}\left(\mathcal{O}\left(\log(p)^4\right)\right)^{1/2} = \mathcal{O}\left(\log(p)^{5/2}n^{-1/2}\right) \\
\frac{1}{n}\left\|\left(\tilde{\mathbf{z}}_j^3\right)^\top \mathbf{x}_{-j}\right\|_\infty &= \mathcal{O}_p\left(\log(p)^{5/2}n^{-1/2}\right) \quad \text{uniformly in } j.
\end{aligned}$$

The last conclusion is a simple application of Markov's inequality. We now turn to the second term to be bounded

$$\begin{aligned}
\frac{1}{n}\left\|\left(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right)^\top \mathbf{x}_{-j}\right\|_\infty &\leq \frac{1}{n}\left\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right\|_1 \|\mathbf{x}_{-j}\|_\infty \\
\|\mathbf{x}_{-j}\|_\infty &= \mathcal{O}_p\left(\sqrt{\log(p)}\right) \quad \text{from the sub-Gaussian assumption.}
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{1}{n}\left\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\right\|_1 &= \frac{1}{n}\left\|\left(\hat{\mathbf{z}}_j - \mathbf{z}_j\right)^3 + 3\mathbf{z}_j \odot \left(\hat{\mathbf{z}}_j - \mathbf{z}_j\right)^2 + 3\mathbf{z}_j^2 \odot \left(\hat{\mathbf{z}}_j - \mathbf{z}_j\right)\right\|_1 \\
&\leq \frac{1}{n}\left\|\left(\hat{\mathbf{z}}_j - \mathbf{z}_j\right)^3\right\|_1 + \frac{3}{n}\left\|\mathbf{z}_j \odot \left(\hat{\mathbf{z}}_j - \mathbf{z}_j\right)^2\right\|_1 + \frac{3}{n}\left\|\mathbf{z}_j^2 \odot \left(\hat{\mathbf{z}}_j - \mathbf{z}_j\right)\right\|_1
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_\infty \Big\|(\hat{\mathbf{z}}_j - \mathbf{z}_j)^2\Big\|_1 + \frac{3}{n}\|\mathbf{z}_j\|_\infty \Big\|(\hat{\mathbf{z}}_j - \mathbf{z}_j)^2\Big\|_1 + \frac{3}{n}\big\|\mathbf{z}_j^2 \odot (\hat{\mathbf{z}}_j - \mathbf{z}_j)\big\|_1 \\
&\leq \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_\infty \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + 3\|\mathbf{z}_j\|_\infty \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + 3\sqrt{\frac{1}{n}\Big\|\mathbf{z}_j^2\Big\|_2^2 \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2} \\
&\leq \|\mathbf{x}_{-j}\|_\infty \big\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\big\|_1 \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + +3\|\mathbf{z}_j\|_\infty \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + 3\sqrt{\frac{1}{n}\Big\|\mathbf{z}_j^2\Big\|_2^2 \frac{1}{n}\|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2} \\
&= \mathcal{O}_p\Big(\sqrt{\log(p)} s_j^2 \lambda_j^3\Big) + \mathcal{O}_p\Big(\sqrt{\log(p)} s_j \lambda_j^2\Big) + \mathcal{O}_p\Big(\sqrt{s_j \lambda_j^2}\Big).
\end{aligned}
$$

In summary,

$$
\frac{1}{n}\big\|\mathbf{x}_{-j}\big(\tilde{\mathbf{z}}_j^3 + (\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3)\big)\big\|_\infty = \mathcal{O}_p\Big(\log(p)^{5/2} n^{-1/2} + \log(p) s_j^2 \lambda_j^3 + \log(p) s_j \lambda_j^2 + \sqrt{\log(p) s_j \lambda_j^2}\Big).
$$

If we choose $\tilde{\lambda}_j$ of this order (as we do in the statement of Lemma 3.8), we have

$$
\Big\|\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big\|_1 = \mathcal{O}_p\Big(\tilde{s}_j \tilde{\lambda}_j\Big) \quad \text{and} \quad \frac{1}{n}\Big\|\mathbf{x}_{-j}\Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)\Big\|_2^2 = \mathcal{O}_p\Big(\tilde{s}_j \tilde{\lambda}_j^2\Big).
$$

For $\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top \mathbf{x}_j$, we use the decomposition

$$
\hat{\tilde{\mathbf{z}}}_j^3 = \hat{\mathbf{z}}_j^3 - \mathbf{x}_{-j}\hat{\tilde{\boldsymbol{\gamma}}}_j = \mathbf{z}_j^3 + (\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3) - \mathbf{x}_{-j}\tilde{\boldsymbol{\gamma}}_j + \mathbf{x}_{-j}\Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big) = \tilde{\mathbf{z}}_j^3 + (\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3) + \mathbf{x}_{-j}\Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big).
$$

Thus,

$$
\begin{aligned}
&\frac{1}{n}\Big|\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top \mathbf{x}_j - \big(\tilde{\mathbf{z}}_j^3\big)^\top \mathbf{x}_j\Big| = \frac{1}{n}\Big|(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3)^\top \mathbf{x}_j + \Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)^\top \mathbf{x}_{-j}^\top \mathbf{x}_j\Big| \\
&\leq \frac{1}{n}\big\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big\|_1 \|\mathbf{x}_j\|_\infty + \frac{1}{n}\Big\|\mathbf{x}_{-j}\Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)\Big\|_2 \|\mathbf{x}_j\|_2 \\
&= \mathcal{O}_p\Big(\log(p) s_j^2 \lambda_j^3 + \log(p) s_j \lambda_j^2 + \sqrt{\log(p) s_j \lambda_j^2}\Big) + \mathcal{O}_p\Big(\sqrt{\tilde{s}_j \tilde{\lambda}_j^2}\Big),
\end{aligned}
$$

which is $\mathcal{o}_p(1)$ by assumption. This leads to

$$
\frac{1}{n}\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top \mathbf{x}_j = \big(\tilde{\mathbf{z}}_j^3\big)^\top \mathbf{x}_j + \mathcal{o}_p(1) = \mathbb{E}\Big[\tilde{Z}_j^3 X_j\Big] + \mathcal{o}_p(1).
$$

The last equality could be derived using the Nemirovski moment inequality in a very similar fashion. For the expectation, we have

$$
\mathbb{E}\Big[\tilde{Z}_j^3 X_j\Big] = \mathbb{E}\Big[\tilde{Z}_j^3\Big(Z_j + \boldsymbol{\gamma}_j^\top \mathbf{X}_{-j}\Big)\Big] = \mathbb{E}\Big[\tilde{Z}_j^3 Z_j\Big] = \mathbb{E}\Big[\Big(Z_j^3 - \tilde{\boldsymbol{\gamma}}_j^\top \mathbf{X}_{-j}\Big) Z_j\Big] = \mathbb{E}\big[Z_j^4\big]
$$

such that the second goal in (3.30) is fulfilled as well. As all these derivations hold uniformly in j, $\Big|\Delta_j^{HOLS}\Big| = \mathcal{o}_p(1)$ implies $\big\|\boldsymbol{\Delta}^{HOLS}\big\| = \mathcal{o}_p(1)$.

3.C.2 Proof of Lemma 3.9

We analyse the error term in (3.28). From the proof of Lemma 3.8 as well as results in van de Geer et al. (2014), we know

$$\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \mathbf{x}_j/n = \mathbb{E}\big[Z_j^4\big] + \mathcal{O}_p(1), \quad \hat{\mathbf{z}}_j^\top \mathbf{x}_j/n = \mathbb{E}\big[Z_j^2\big] + \mathcal{O}_p(1) \quad \text{and} \quad \|\hat{\mathbf{z}}_j\|_2^2/n = \mathbb{E}\big[Z_j^2\big] + \mathcal{O}_p(1).$$

For the remaining terms in $\hat{\mathbf{v}}_j^\top \hat{\mathbf{v}}_j/n$, we want to ensure

$$\frac{1}{n}\left\|\hat{\tilde{\mathbf{z}}}_j^3\right\|_2^2 = \mathbb{E}\Big[\tilde{Z}_j^6\Big] + \mathcal{O}_p(1) \quad \text{and} \quad \frac{1}{n}\left(\hat{\tilde{\mathbf{z}}}_j^3\right)^\top \hat{\mathbf{z}}_j = \left\|\hat{\mathbf{z}}_j^2\right\|_2^2/n = \big[Z_j^4\big] + \mathcal{O}_p(1). \tag{3.31}$$

Using the Nemirovski equation in a similar fashion as before, we know

$$\max_j \left| \frac{1}{n}\sum_{i=1}^n z_{ij}^r - \mathbb{E}\big[Z_j^r\big] \right| = \mathcal{O}_p\left(\frac{\log(p)^{(r+1)/2}}{n^{1/2}}\right).$$

We assume this to be $\mathcal{O}_p(1)\ \forall r \leq 10$ and even $o_p(1)\ \forall r \leq 6$ (which is implied by the first condition). We look at some intermediary results. Each difference is $o_p(1)$ using the sparsity assumptions.

$$\begin{aligned}
\frac{1}{n}\big\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^2 - \mathbf{z}_j^4\big\|_1 &= \frac{1}{n}\big\|\mathbf{z}_j^2 \odot \big(\hat{\mathbf{z}}_j^2 - \mathbf{z}_j^2\big)\big\|_1 = \frac{1}{n}\Big\|\mathbf{z}_j^2 \odot \Big((\hat{\mathbf{z}}_j - \mathbf{z}_j)^2 + 2\mathbf{z}_j \odot (\hat{\mathbf{z}}_j - \mathbf{z}_j)\Big)\Big\|_1 \\
&\leq \frac{1}{n}\big\|\mathbf{z}_j^2\big\|_\infty \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + \frac{2}{n}\big\|\mathbf{z}_j^3\big\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2 \\
&= \mathcal{O}_p\big(\log(p) s_j \lambda_j^2\big) + \mathcal{O}_p\Big(\sqrt{s_j\lambda_j^2}\Big).
\end{aligned}$$

This implies $\frac{1}{n}\Big\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^2\Big\|_1 = \frac{1}{n}\|\mathbf{z}_j \odot \hat{\mathbf{z}}_j\|_2^2 = \mathcal{O}_p(1)$. With this, we can refine our result in a stepwise fashion:

$$\begin{aligned}
\frac{1}{n}\big\|\mathbf{z}_j^4 \odot \hat{\mathbf{z}}_j^2 - \mathbf{z}_j^6\big\|_1 &= \frac{1}{n}\big\|\mathbf{z}_j^4 \odot \big(\hat{\mathbf{z}}_j^2 - \mathbf{z}_j^2\big)\big\|_1 = \frac{1}{n}\Big\|\mathbf{z}_j^4 \odot \Big((\hat{\mathbf{z}}_j - \mathbf{z}_j)^2 + 2\mathbf{z}_j \odot (\hat{\mathbf{z}}_j - \mathbf{z}_j)\Big)\Big\|_1 \\
&\leq \frac{1}{n}\big\|\mathbf{z}_j^4\big\|_\infty \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + \frac{2}{n}\big\|\mathbf{z}_j^5\big\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2 \\
&= \mathcal{O}_p\Big(\log(p)^2 s_j \lambda_j^2\Big) + \mathcal{O}_p\Big(\sqrt{s_j\lambda_j^2}\Big) \\
\text{such that } \frac{1}{n}\big\|\mathbf{z}_j^4 \odot \hat{\mathbf{z}}_j^2\big\|_1 &= \frac{1}{n}\big\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j\big\|_2^2 = \mathcal{O}_p(1).
\end{aligned}$$

$$\begin{aligned}
\frac{1}{n}\big\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^4 - \mathbf{z}_j^4 \odot \hat{\mathbf{z}}_j^2\big\|_1 &= \frac{1}{n}\big\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^2 \odot \big(\hat{\mathbf{z}}_j^2 - \mathbf{z}_j^2\big)\big\|_1 \\
&= \frac{1}{n}\Big\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^2 \odot \Big((\hat{\mathbf{z}}_j - \mathbf{z}_j)^2 + 2\mathbf{z}_j \odot (\hat{\mathbf{z}}_j - \mathbf{z}_j)\Big)\Big\|_1
\end{aligned}$$

$$\leq \frac{1}{n}\|\mathbf{z}_j^2\|_\infty \|\hat{\mathbf{z}}_j^2\|_\infty \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + \frac{2}{n}\|\mathbf{z}_j\|_\infty \|\hat{\mathbf{z}}_j\|_\infty \|\mathbf{z}_j^2 \hat{\mathbf{z}}_j\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2$$
$$= \mathcal{O}_p\Big(\log(p)^2\big(1 + s_j^2\lambda_j^2\big)s_j\lambda_j^2\Big) + \mathcal{O}_p\Big(\log(p)(1 + s_j\lambda_j)\sqrt{s_j\lambda_j^2}\Big)$$
such that $\frac{1}{n}\|\mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^4\|_1 = \frac{1}{n}\|\mathbf{z}_j \odot \hat{\mathbf{z}}_j^2\|_2^2 = \mathcal{O}_p(1)$.

$$\frac{1}{n}\|\hat{\mathbf{z}}_j^6 - \mathbf{z}_j^2 \odot \hat{\mathbf{z}}_j^4\|_1 = \frac{1}{n}\|\hat{\mathbf{z}}_j^4 \odot \big(\hat{\mathbf{z}}_j^2 - \mathbf{z}_j^2\big)\|_1 = \frac{1}{n}\Big\|\hat{\mathbf{z}}_j^4 \odot \Big((\hat{\mathbf{z}}_j - \mathbf{z}_j)^2 + 2\mathbf{z}_j \odot (\hat{\mathbf{z}}_j - \mathbf{z}_j)\Big)\Big\|_1$$
$$\leq \frac{1}{n}\|\hat{\mathbf{z}}_j^4\|_\infty \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2^2 + \frac{2}{n}\|\hat{\mathbf{z}}_j^2\|_\infty \|\mathbf{z}_j \hat{\mathbf{z}}_j^2\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2$$
$$= \mathcal{O}_p\Big(\log(p)^2\big(1 + s_j^4\lambda_j^4\big)s_j\lambda_j^2\Big) + \mathcal{O}_p\Big(\log(p)\big(1 + s_j^2\lambda_j^2\big)\sqrt{s_j\lambda_j^2}\Big).$$

Finally, it follows

$$\frac{1}{n}\|\mathbf{z}_j^3 \odot \hat{\mathbf{z}}_j^3 - \mathbf{z}_j^4 \odot \hat{\mathbf{z}}_j^2\|_1 = \frac{1}{n}\|\mathbf{z}_j^3 \odot \hat{\mathbf{z}}_j^2 \odot (\hat{\mathbf{z}}_j - \mathbf{z}_j)\|_1 = \frac{1}{n}\|\mathbf{z}_j^2\|_\infty \|\mathbf{z}_j \hat{\mathbf{z}}_j^2\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2$$
$$= \mathcal{O}_p\Big(\log(p)\sqrt{s_j\lambda_j^2}\Big) \text{ such that}$$
$$\frac{1}{n}\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\|_2^2 = \frac{1}{n}\|\hat{\mathbf{z}}_j^6 + \mathbf{z}_j^6 - 2\hat{\mathbf{z}}_j^3 \odot \mathbf{z}_j^3\|_1 \leq \frac{1}{n}\|\hat{\mathbf{z}}_j^6 - \hat{\mathbf{z}}_j^3 \odot \mathbf{z}_j^3\|_1 + \frac{1}{n}\|\mathbf{z}_j^6 - \hat{\mathbf{z}}_j^3 \odot \mathbf{z}_j^3\|_1$$
$$= \mathcal{O}_p\Big(\log(p)^2\big(1 + s_j^4\lambda_j^4\big)s_j\lambda_j^2\Big) + \mathcal{O}_p\Big(\log(p)\big(1 + s_j^2\lambda_j^2\big)\sqrt{s_j\lambda_j^2}\Big)$$
$$= o_p(1),$$

This can now be applied to find the desired convergence.

$$\frac{1}{n}\Big|\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top \hat{\mathbf{z}}_j - \big(\tilde{\mathbf{z}}_j^3\big)^\top \mathbf{z}_j\Big| = \frac{1}{n}\Big|\big(\tilde{\mathbf{z}}_j^3\big)^\top(\hat{\mathbf{z}}_j - \mathbf{z}_j) + \big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)^\top \mathbf{z}_j +$$
$$\big(\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\big)^\top(\hat{\mathbf{z}}_j - \mathbf{z}_j) + \Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)^\top \mathbf{x}_{-j}^\top \mathbf{z}_j + \Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)^\top \mathbf{x}_{-j}^\top(\hat{\mathbf{z}}_j - \mathbf{z}_j)\Big| \leq$$
$$\frac{1}{n}\|\tilde{\mathbf{z}}_j^3\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2 + \frac{1}{n}\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\|_2 \|\mathbf{z}_j\|_2 + \frac{1}{n}\|\hat{\mathbf{z}}_j^3 - \mathbf{z}_j^3\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2 +$$
$$\frac{1}{n}\|\mathbf{z}_j\|_2 \Big\|\mathbf{x}_{-j}\Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)\Big\|_2 + \frac{1}{n}\Big\|\mathbf{x}_{-j}\Big(\tilde{\boldsymbol{\gamma}}_j - \hat{\tilde{\boldsymbol{\gamma}}}_j\Big)\Big\|_2 \|\hat{\mathbf{z}}_j - \mathbf{z}_j\|_2$$

All these terms have already been bounded. Thus, we do not need any further assumptions to claim

$$\frac{1}{n}\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top \hat{\mathbf{z}}_j = \frac{1}{n}\ \big(\tilde{\mathbf{z}}_j^3\big)^\top \mathbf{z}_j + o_p(1) = \mathbb{E}\Big[\tilde{Z}_j^3 Z_j\Big] + o_p(1) = \mathbb{E}\big[Z_j^4\big] + o_p(1).$$

We turn to the final term in the error scaling

$$\frac{1}{n}\Big|\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top \Big(\hat{\tilde{\mathbf{z}}}_j^3\Big) - \big(\tilde{\mathbf{z}}_j^3\big)^\top \big(\tilde{\mathbf{z}}_j^3\big)\Big| =$$

$$\frac{1}{n}\Big|2\big(\tilde{\mathbf{z}}_j^3\big)^\top\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)+2\big(\tilde{\mathbf{z}}_j^3\big)^\top\mathbf{x}_{-j}\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)+\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)^\top\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)+$$
$$2\big(\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big)\mathbf{x}_{-j}\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)+\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)^\top\mathbf{x}_{-j}^\top\mathbf{x}_{-j}\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)\Big|\leq$$
$$\frac{2}{n}\big\|\tilde{\mathbf{z}}_j^3\big\|_2\big\|\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big\|_2+\frac{2}{n}\big\|\tilde{\mathbf{z}}_j^3\big\|_2\Big\|\mathbf{x}_{-j}\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)\Big\|_2+\frac{1}{n}\big\|\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big\|_2^2+$$
$$\frac{2}{n}\big\|\hat{\mathbf{z}}_j^3-\mathbf{z}_j^3\big\|_2\Big\|\mathbf{x}_{-j}\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)\Big\|_2+\frac{1}{n}\Big\|\mathbf{x}_{-j}\Big(\tilde{\gamma}_j-\hat{\tilde{\gamma}}_j\Big)\Big\|_2^2.$$

Again, these are all terms that we have seen before such that we do not need any additional assumptions to claim

$$\frac{1}{n}\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)^\top\Big(\hat{\tilde{\mathbf{z}}}_j^3\Big)=\frac{1}{n}\big(\tilde{\mathbf{z}}_j^3\big)^\top\big(\tilde{\mathbf{z}}_j^3\big)+\mathcal{o}_p(1)=\mathbb{E}\Big[\Big(\tilde{Z}_j^3\Big)^2\Big]+\mathcal{o}_p(1).$$

Thus, we have shown convergence for all terms in $\hat{\mathbf{v}}_j^\top\hat{\mathbf{v}}_j/n$. Finally, note that

$$\begin{aligned}\mathbb{E}\big[V_j^2\big]&=\mathbb{E}\left[\left(\frac{\tilde{Z}_j^3}{\mathbb{E}\Big[Z_j^4\Big]}-\frac{Z_j}{\mathbb{E}\Big[Z_j^2\Big]}\right)^2\right]=\frac{\mathbb{E}\Big[\Big(\tilde{Z}_j^3\Big)^2\Big]}{\mathbb{E}\Big[Z_j^4\Big]^2}-\frac{2\mathbb{E}\Big[\tilde{Z}_j^3Z_j\Big]}{\mathbb{E}\Big[Z_j^4\Big]\mathbb{E}\Big[Z_j^2\Big]}+\frac{\mathbb{E}\Big[Z_j^2\Big]}{\mathbb{E}\Big[Z_j^2\Big]^2}\\&=\frac{\mathbb{E}\Big[\Big(\tilde{Z}_j^3\Big)^2\Big]}{\mathbb{E}\Big[Z_j^4\Big]^2}-\frac{1}{\mathbb{E}\Big[Z_j^2\Big]}\end{aligned}$$

such that convergence is towards $\mathbb{E}\Big[V_j^2\Big]$ as claimed.

3.C.3 Proof of Theorem 3.2

With Lemmata 3.8 and 3.9, we have already established that the bias terms vanish and the denominator converges. It remains to look at $\hat{\mathbf{v}}_j^\top\boldsymbol{\epsilon}/\sqrt{n}$.

$$\begin{aligned}\mathbb{E}\Big[(\hat{\mathbf{v}}_j-\mathbf{v}_j)^\top\boldsymbol{\epsilon}/\sqrt{n}\Big]&=\mathbb{E}\Big[\mathbb{E}\Big[(\hat{\mathbf{v}}_j-\mathbf{v}_j)^\top\boldsymbol{\epsilon}/\sqrt{n}|\mathbf{x}\Big]\Big]=\mathbb{E}\Big[(\hat{\mathbf{v}}_j-\mathbf{v}_j)^\top\mathbb{E}\big[\boldsymbol{\epsilon}/\sqrt{n}|\mathbf{x}\big]\Big]=0\quad\text{and}\\\operatorname{Var}\Big((\hat{\mathbf{v}}_j-\mathbf{v}_j)^\top\boldsymbol{\epsilon}/\sqrt{n}\Big)&=\mathbb{E}\Big[\operatorname{Var}\Big((\hat{\mathbf{v}}_j-\mathbf{v}_j)^\top\boldsymbol{\epsilon}/\sqrt{n}|\mathbf{x}\Big)\Big]+\operatorname{Var}\Big(\mathbb{E}\Big[(\hat{\mathbf{v}}_j-\mathbf{v}_j)^\top\boldsymbol{\epsilon}/\sqrt{n}|\mathbf{x}\Big]\Big)\\&=\sigma^2\mathbb{E}\Big[\|\hat{\mathbf{v}}_j-\mathbf{v}_j\|_2^2/n\Big]+0=\mathcal{o}(1)\end{aligned}$$

The last equality uses the convergence rates from the proof of Lemma 3.9. By Chebyshev's inequality and the CLT

$$\hat{\mathbf{v}}_j^\top\boldsymbol{\epsilon}/\sqrt{n}\xrightarrow{\mathbb{P}}\mathbf{v}_j^\top\boldsymbol{\epsilon}/\sqrt{n}\xrightarrow{\mathbb{D}}\mathcal{N}\big(0,\sigma^2\mathbb{E}\big[V_j^2\big]\big)$$

By Slutsky's theorem, we can replace σ^2 with a consistent estimate such that the theorem's statement follows.

Chapter 4

Assessing the overall and partial causal well-specification of nonlinear additive noise models

Christoph Schultheiss and Peter Bühlmann
Journal of Machine Learning Research 25, (159): 1-41.

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.

4.1 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. (2008a); 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. (2019b). 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. (2019b) 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. (2024).

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 4.2.4 which illustrates some gains over the approach by Maeda and Shimizu (2021) by exploiting conditional independence.

4.2 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 4.3 and 4.6.1.

In Section 4.2.1, we introduce the causal model, our notation, and the most important background concepts from the causality literature. Section 4.2.2 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 4.2.3 and 4.2.4.

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, $\mathfrak{C}$,

$$\mathfrak{C}: \quad Z_j \leftarrow f_j\big(\mathbf{Z}_{\mathrm{PA}(j)}, \xi_j\big) \ \forall j \in \{1, \ldots, q\}. \tag{4.1}$$

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 (4.1), 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

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

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 $\mathbf{H}_{\mathrm{PA}(Y)}$. Therefore, $\mathbf{H}_{\mathrm{PA}(Y)}$ always has dimensionality of at least one assuming Y is not deterministic in $\mathbf{X}$. In our SCM (4.1), we then have

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

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

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

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

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

As the Markov boundary is defined within only the observed covariates $\mathbf{X}$, in a structure $H_1 \to H_2 \to X_1$, 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 4.2.4. It could also be all of $\mathbf{X}$ or empty.

We use causal do-notation to denote interventions. Conditioning on, e.g., $\mathrm{do}(Z_j \leftarrow z_j)$ means that we assume a variation of the SCM $\mathfrak{C}$ (4.1) 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, $\mathrm{do}(\mathbf{Z}_S \leftarrow \mathbf{z}_S)$ 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., $P_Y^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\mathrm{do}(\mathbf{X}\leftarrow\mathbf{x}')}$ denotes the counterfactual distribution of Y in the SCM $\mathfrak{C}$ where $\mathbf{Z} = \mathbf{z}$ is observed and the counterfactual intervention is $\mathbf{X} \leftarrow \mathbf{x}'$.

4.2.2 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 4.1. Detailed mathematical definitions, assumptions, and results are given in the subsequent sections.

The additive noise model (ANM) has the following structure

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

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 (4.1) $\forall k \in N$

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 ((A4.2) later) and are conditionally independent of these given the remaining covariates ((A4.1) 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 (4.1) $\forall k \in N$ 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 4.1.

Importantly, our results do not assume any model apart from the SCM in (4.1). 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 (4.1), 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.

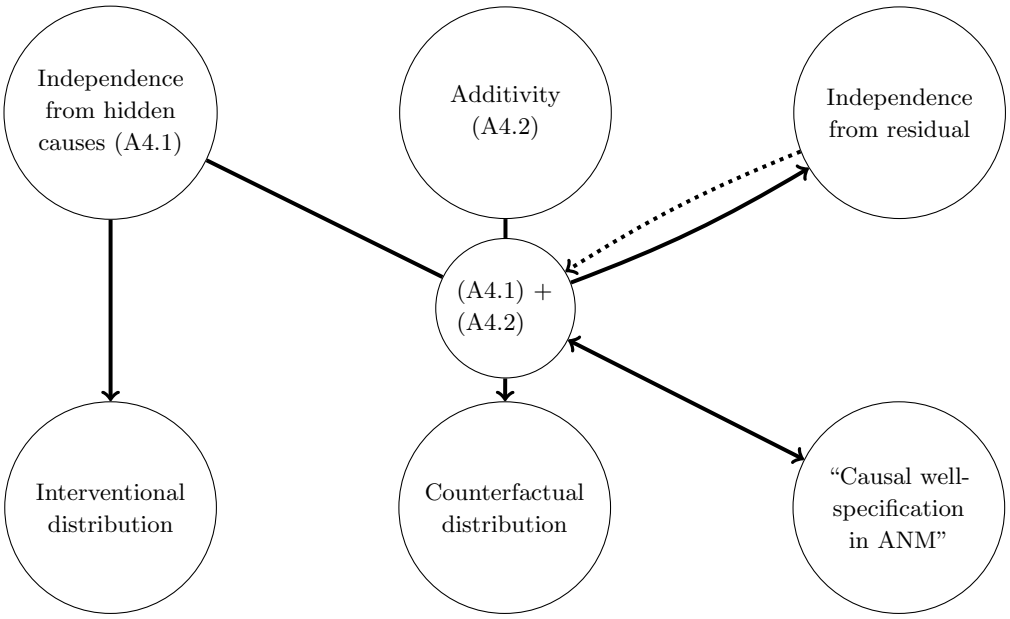


Figure 4.1: Interconnection between assumptions, causal interpretations, and the testable proxy. Directed edges denote implications, the plus means that both assumptions hold simultaneously, the bidirected edge denotes equivalence per definition, and the dotted edge denotes implication up to pathological cases, i.e., a proxy.

4.2.3 Global well-specification

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

$$Y \leftarrow f_{\mathbf{X}Y}\big(\mathbf{X}_{\mathrm{PA}(Y)}\big) + f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}\big), \quad \text{where} \quad \mathbf{X} \perp \mathbf{H}_{\mathrm{PA}(Y)}. \tag{4.2}$$

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

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

where $\stackrel{d}{=}$ 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 (4.2) 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,

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

The conditions in (4.2) additionally imply the following global null hypothesis that we aim to check first.

$$H_0: \quad \mathcal{E} \perp \mathbf{X}, \quad \text{where} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}]. \tag{4.3}$$

Note the subtle difference between (4.2) and (4.3). The conditions in (4.2) are sufficient to fulfill (4.3) 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 (4.2) 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, (4.3) is a useful proxy for (4.2), 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. (2008a) and Peters et al. (2014).

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

4.2.4 Local well-specification

If the conditions (4.2) are partially violated it might still be possible to correctly understand the causal effect for *some* of the predictors $\mathbf{X}_U$ where $U \subseteq \{1,\dots,p\}$. 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.

(A4.1) The covariates in $\mathbf{X}_{-U}$ form a Markov blanket of $\mathbf{H}_{\mathrm{PA}(Y)}$, i.e., $\mathbf{H}_{\mathrm{PA}(Y)} \perp \mathbf{X}_U|\mathbf{X}_{-U}$.

(A4.2) $Y \leftarrow f_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big) + f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}, \mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)$, i.e., 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 4.2 containing two examples of DAGs with hidden parents. On the left, the set $U = \{2\}$ fulfils (A4.1). On the right, $U = \{1\}$ fulfils it. Correctness of (A4.2) depends on the

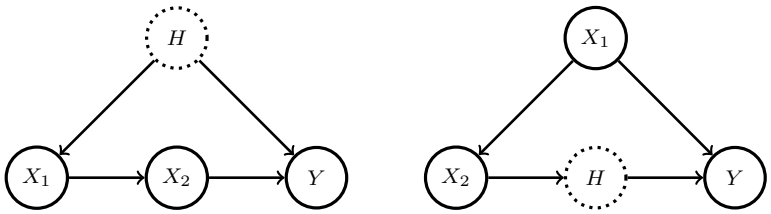


Figure 4.2: Left: Structure with a hidden confounder. Right: Structure with a hidden mediator.

structural assignment for Y. E.g.,

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

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 (A4.1). 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 (A4.1): 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 (B4.1) is violated; see below.

We impose no constraints onto $f_{\mathbf{X}_U Y}(\cdot)$ and $f_{\mathbf{H}Y}(\cdot)$ in (A4.2). Hence, the first summand could be zero, which is the case for $U \cap \mathrm{PA}(Y) = \emptyset$.

(A4.1) ensures that

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

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 (A4.2), we get two implications under an additional technical assumption.

(B4.1) Let $\{A, B, C\}$ be disjoint subsets of $\{1, \ldots, q\}$ in model (4.1). Then,

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

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 (B4.1) would be the following SCM with continuous and univariate components Z_A, Z_B, and Z_C

$$Z_C \leftarrow \xi_C,$$
$$Z_B \leftarrow Z_C + \xi_B,$$

$$Z_A \leftarrow Z_C + \xi_A + \mathbb{1}_{\{Z_B=0\}}.$$

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 4.1. Assume the model (4.1) with (B4.1). Let $\mathbf{X}_U$ be a set of covariates fulfilling (A4.1) and (A4.2), then

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

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

$$H_{0,U}: \quad \mathcal{E} \perp \mathbf{X}_U | \mathbf{X}_{-U}, \quad \text{with} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}]. \tag{4.4}$$

The first implication means that in the counterfactual, where we can change $\mathbf{X}_U$ without changing $\mathbf{X}_{-U}$ or ξ_k in (4.1) $\forall k \notin M$, 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 (A4.1) and (B4.1), 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 4.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

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

which are equivalent if both are defined as discussed above. However, this is not estimable outside the data support. $H_{0,U}$ (4.4) 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 (A4.1) 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 (A4.1) and (A4.2) 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 (A4.1) and (A4.2). Again, a multivariate Gaussian distribution is an example where (4.4) holds regardless of (A4.1). However, for other data generating distributions, we consider (4.4) to be a good proxy to see whether (A4.1) and (A4.2) might hold.

Of most interest are the sets

$$W \in \underset{U:H_{0,U} \text{ is true}}{\arg\max} |U|. \tag{4.5}$$

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).

(B4.2) $\mathcal{E} \perp \mathbf{X}_A | \mathbf{X}_B, \mathbf{X}_C \quad \text{and} \quad \mathcal{E} \perp \mathbf{X}_B | \mathbf{X}_A, \mathbf{X}_C \implies \mathcal{E} \perp \mathbf{X}_A, \mathbf{X}_B | \mathbf{X}_C$
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

$$\begin{aligned} H_{0,j}: \quad & \mathcal{E} \perp X_j | \mathbf{X}_{-j}, \quad \text{where} \quad \mathcal{E} = Y - \mathbb{E}[Y|\mathbf{X}] \\ & \tilde{W} = \{j : H_{0,j} \text{ is true}\}. \end{aligned} \tag{4.6}$$

We can relate this to the Markov boundary.

Theorem 4.2. Assume the model (4.1). Let W be any set as in (4.5) and $\tilde{W}$ as in (4.6). Then, $W \subseteq \tilde{W}$. If the intersection property (B4.2) holds, $W = \tilde{W}$, and W is unique.

In general, (A4.1) and (A4.2) do not imply that the ANM (4.2) 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 (4.3).

Recall the examples in Figure 4.2. In the left structure, $W = \{2\}$ if (A4.2) 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 (A4.2) 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 4.2 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.

4.3 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 4.3.3.

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

4.3.1 Making use of FOCI (Feature Ordering by Conditional Independence)

One estimation strategy would be to test the hypotheses in (4.6) 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 $\Big(\{1, \dots, p\} \setminus \hat{S}\Big) \subseteq W$.

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 4.3.2. 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 4.3.2.1. Furthermore, we show a new result for transforming the data before applying FOCI; see Proposition 4.2. Finally, we suggest an algorithm that yields more stable estimates in Section 4.3.3.

We provide now some background on FOCI by focusing on its main concepts. The precise definitions can be found in Appendix 4.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.,

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

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 $\{1, \dots, p\} \setminus W$ with high probability. However, there is in general no guarantee against superfluous inclusion to $\hat{S}$ and we get $\left(\{1, \dots, p\} \setminus \hat{S}\right) \subseteq W$.

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 4.A.3 here. FOCI greedily increases the set of predictors to maximize Q.

Proposition 4.1. Let $S \subseteq \{1, \dots, p\}$. 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.

Algorithm 4.1 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 (4.4) 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 4.2 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 (4.4) 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 $\hat{S}$

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 4.2 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 4.3.3.

4.3.2 Asymptotic results

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

(B4.3) $|\hat{\epsilon}_i - \epsilon_i| = \mathcal{O}_p(1)$.

(B4.4) $\mathcal{E}$ is a continuous random variable.

(B4.5) $\nexists S \subseteq \{1, \ldots, p\}$ 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 (B4.3) 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 (B4.4) for the proofs, we provide a simple example in Section 4.3.2.1 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 $t \in \mathbb{R}$,

$$\begin{aligned} &\big| P(g(\mathcal{E}) \geq t | \mathbf{X} = \mathbf{x}_S) - P(g(\mathcal{E}) \geq t | \mathbf{X} = \mathbf{x}'_S) \big| \\ &\leq C(1 + \|\mathbf{x}_S\|^\beta + \|\mathbf{x}'_S\|^\beta) \|\mathbf{x}_S - \mathbf{x}'_S\|. \end{aligned}$$

(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 4.3. 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 (B4.2) - (B4.5). Let $\hat{S}$ be the output of Algorithm 4.2. There are positive real numbers L_1, L_2 and L_3 that do not depend on the sample size such that

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

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 4.4. Assume the conditions of Theorem 4.3. Let $\hat{S}$ be the output of Algorithm 4.1. Then,

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

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.

4.3.2.1 Discrete $\mathcal{E}$

For our results, we invoked assumption (B4.4), 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 4.A.3.

Proposition 4.2. 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 (B4.3), 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 4.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.

4.3.3 Practical algorithm

Although we can consistently find a Markov blanket (but not necessarily the minimal Markov boundary) using Algorithms 4.1 or 4.2 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 4.2 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 (4.3). 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

Algorithm 4.3 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

$n_j = 0 \; \forall j = 1, \ldots, p$
for $b = 1$ to B **do**
  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 $\left(\mathbf{x}^{(2)}, \mathbf{y}^{(2)}\right)$
  Get an estimate $\hat{f}(\mathbf{X})$ for $\mathbb{E}[Y|\mathbf{X}]$ using a certain regressor on the data $\left(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}\right)$
  Estimate the residuals as $\hat{\boldsymbol{\epsilon}}^{(2)} = \mathbf{y}^{(2)} - \hat{f}\left(\mathbf{x}^{(2)}\right)$
  Apply the HSIC test to the data $\left(\hat{\boldsymbol{\epsilon}}^{(2)}, \mathbf{x}^{(2)}\right)$ to get the p-value p^b.
  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 $\hat{S}^b$
  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 $\hat{S}^{B+b}$
Combine the p-values $p^1, \ldots p^{2B}$ (Meinshausen et al., 2009) and get the model p-value p^0.
if $p^0 > \alpha$ **then**
  $\hat{W} = \{1, \ldots, p\}$
else
  $\hat{W} = \emptyset$
  for $j = 1$ to p **do**
    $n_j = \sum_{b=1}^{2B} j \in \hat{S}^b$
  $\bar{n} = \sum_{j=1}^{p} n_j / p$
  $n^{\min} = \min_{j: n_j \geq \bar{n}} n_j$
  for $j = 1$ to p **do**
    if $n_j < \bar{n}$ and $proportion.test\left(n_j, n^{min}, 2B\right) \leq \tilde{\alpha}$ **then**
      $\hat{W} = \hat{W} \cup j$

necessary variables are selected by FOCI in almost every split, see Theorem 4.3, 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 $\mathbb{P}\Big(j \in \hat{W}\Big)$ 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 4.3.

4.4 Simulation example

We evaluate the method on a simple SCM represented by the DAG in Figure 4.3. We let the causal effects be non-monotonic functions. As discussed in Section 5.3, 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 4.3 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

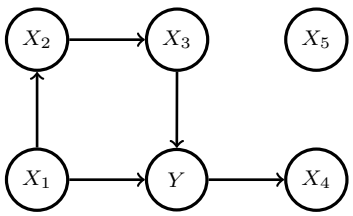


Figure 4.3: DAG representing the SCM in the simulation.

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) $\hat{\mathbb{P}}\Big(j \in \hat{W} | j \notin W\Big)$ and the true positive rate (TPR) $\hat{\mathbb{P}}\Big(j \in \hat{W} | j \in W\Big)$ 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 4.4.

For $n = 10^2$, a low FPR is not attainable because the p-values for (4.3) are not reliably small enough, and Algorithm 4.3 often terminates before considering the individual covariates. However, even for this low sample size, we get a performance that is clearly better than random guessing. For large sample sizes, the FPR becomes very low which is in agreement with Theorem 4.3. 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 4.2 (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 dataset can lead to a performance boost.

We also evaluate the testing of H_0 (4.3). For this, we show the empirical cumulative distribution function of the obtained p-values in the middle of Figure 4.4. 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 $\alpha = 0.05$, 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

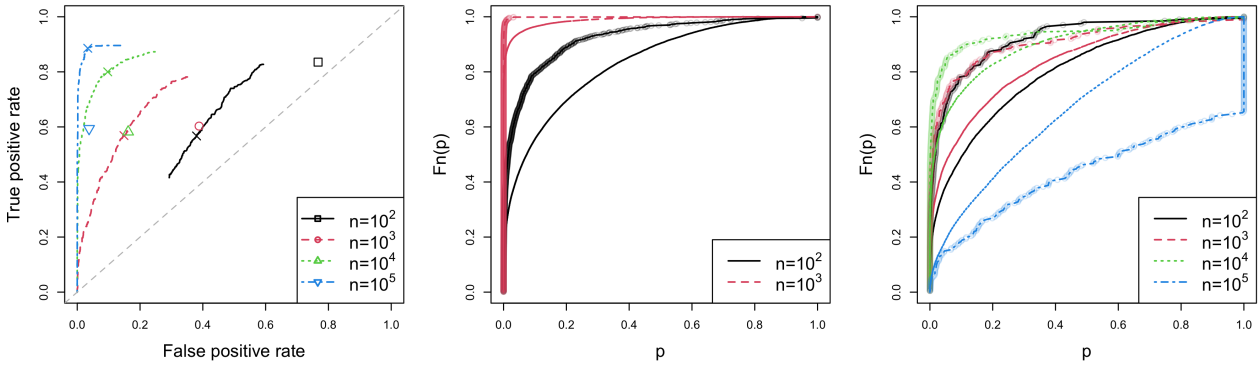


Figure 4.4: The results are based on 100 simulation runs. On the left: False positive rate versus true positive rate obtained with Algorithm 4.3 for varying $\tilde{\alpha}$ and $\alpha = 0.05$. The crosses correspond to $\tilde{\alpha} = 0.01$. The other symbols describe the performance of Algorithm 4.2. 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.

well-specified models. This is shown on the right side of Figure 4.4. 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 (4.3) 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.

4.5 Real data analysis

We consider the K562 dataset provided by Replogle et al. (2022). We follow the preprocessing in the benchmark of Chevalley et al. (2023). Then, the dataset 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 4.3 with $B = 25$ splits using `xgboost` for regression. There is a strong indication against the global null hypothesis (4.3) 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 4.3 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.,

$$RB^{X_j \to Y} = \frac{\left|\sum_{i \in \mathcal{D}_j} y_i - \hat{f}(\mathbf{x}_i)\right| / |\mathcal{D}_j|}{\sum_{i \in \mathcal{D}_\mathcal{O}} y_i / |\mathcal{D}_\mathcal{O}|}, \tag{4.7}$$

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 (4.7) 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 4.2.4. 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

Target Y	Predictor X_j	Mann-Whitney U test	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 4.1: Application to the K562 dataset 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 4.3, 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 $RB^{X_j \to Y}$ (4.7) when using the model fit on observational data to predict the target in the dataset where the predictor is knocked down.

corresponds to the last row-box in Table 4.1.

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 4.1, 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 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.

4.6 Location-scale noise models

A simple extension of model (4.2), 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 $g_{\mathbf{X}Y}\big(\mathbf{X}_{\mathrm{PA}(Y)}\big)$ 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, see also Chapter 6. In analogy to (4.2), we call the LSNM causally well-specified if

$$Y \leftarrow f_{\mathbf{X}Y}\big(\mathbf{X}_{\mathrm{PA}(Y)}\big) + g_{\mathbf{X}Y}\big(\mathbf{X}_{\mathrm{PA}(Y)}\big) f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)}\big), \quad \text{where} \quad \mathbf{H}_{\mathrm{PA}(Y)} \perp \mathbf{X}. \tag{4.8}$$

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

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

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 (4.8), we have the natural proxy

$$H_0: \quad \mathcal{E} \perp \mathbf{X}, \quad \text{where} \quad \mathcal{E} = \frac{Y - \mathbb{E}[Y|\mathbf{X}]}{\sqrt{\text{Var}(Y|\mathbf{X})}} \tag{4.9}$$

since under (4.8) we have that $\mathcal{E} = f_{\mathbf{H}Y}\big(\mathbf{H}_{\text{PA}(Y)}\big)$ and hence H_0 in (4.9) holds.

In case of model misspecification, we can consider the per-covariate causal well-specification. Condition (A4.1) remains the same for the LSNM, (A4.2) can be replaced by a weaker version for this more flexible causal model:

(A4.2*) $Y \leftarrow f_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\text{PA}(Y)\setminus U}\big) + g_{\mathbf{X}_U Y}\big(\mathbf{X}_U, \mathbf{X}_{\text{PA}(Y)\setminus U}\big) f_{\mathbf{H}Y}\big(\mathbf{H}_{\text{PA}(Y)}, \mathbf{X}_{\text{PA}(Y)\setminus U}\big)$, 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 4.5. Assume the model (4.1) with (B4.1). Let $\mathbf{X}_U$ be a set of covariates fulfilling (A4.1) and (A4.2*), then

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

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

$$H_{0,U}: \quad \mathcal{E} \perp \mathbf{X}_U | \mathbf{X}_{-U}, \quad \text{with} \quad \mathcal{E} = \frac{Y - \mathbb{E}[Y|\mathbf{X}]}{\sqrt{\text{Var}(Y|\mathbf{X})}}. \tag{4.10}$$

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 4.2.4 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.

4.6.1 Asymptotic results

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

We consider variations of Algorithms 4.1 and 4.2 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 $f_2(\mathbf{X}) \coloneqq \mathbb{E}\big[Y^2|\mathbf{X}\big]$ using certain regressors on the data $(\mathbf{x}, \mathbf{y})$; see the notation in Section 4.3. 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

(B4.6) $\left|f_1(\mathbf{x}_i) - \hat{f}_1(\mathbf{x}_i)\right| = \mathcal{O}_p(1)$.

(B4.7) $\left|f_2(\mathbf{x}_i) - \hat{f}_2(\mathbf{x}_i)\right| = \mathcal{O}_p(1)$.

(B4.8) $\mathbb{P}\big(f_2(\mathbf{x}_i) - f_1^2(\mathbf{x}_i) > 0\big) = 1$.

In Assumptions (B4.6) and (B4.7) the probability is over both, the function estimates and the new data point. Assumption (B4.8) implies that Y is almost surely not deterministic in $\mathbf{X}$.

Theorem 4.6. 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 (B4.2) and (B4.4) - (B4.8). Let $\hat{S}$ be the output of Algorithm 4.2 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 4.1 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 (B4.6) - (B4.8) 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.

4.6.2 Simulation example

We consider a simple example with two observed predictors and one hidden confounder as shown in Figure 4.5. We let

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

such that (A4.2*) 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.

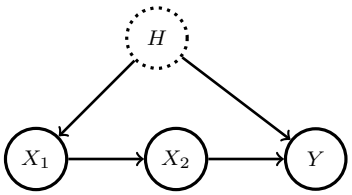


Figure 4.5: DAG representing the SCM in the simulation fitting LSNM.

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 4.4.

In Figure 4.6, we show the same performance metrics as in Figure 4.4. 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 $\hat{W}$.

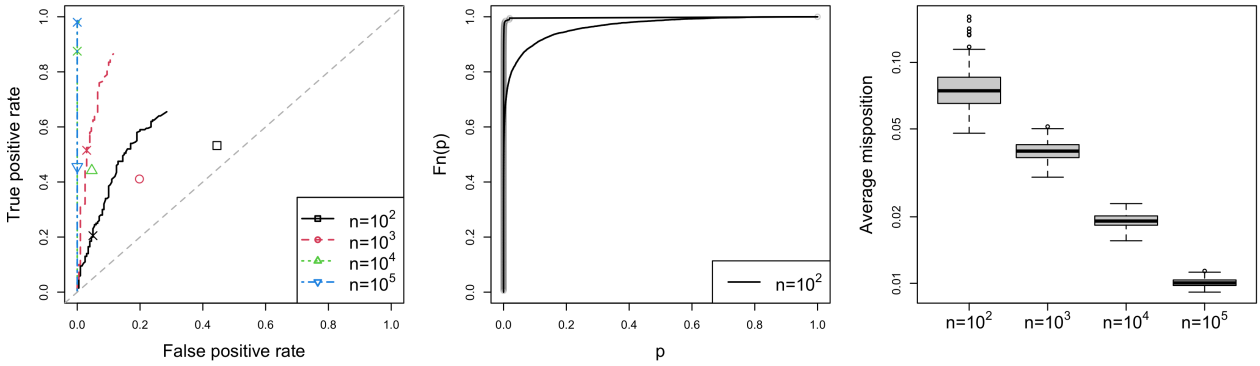


Figure 4.6: The results are based on 200 simulation runs. On the left: False positive rate versus true positive rate obtained with Algorithm 4.3 adjusted to LSNM for varying $\tilde{\alpha}$ and $\alpha = 0.05$. The crosses correspond to $\tilde{\alpha} = 0.01$. The other symbols describe the performance of Algorithm 4.2 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 (4.11) of the estimated residuals with respect to the true residuals.

The global test works well already for 10^2 samples. After aggregation over the different splits, H_0 in (4.9) 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 $\hat{\boldsymbol{\epsilon}}$ than for the true $\boldsymbol{\epsilon}$.

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

$$AMP = \frac{1}{n^2} \sum_{i=1}^{n} \left| \sum_{l=1}^{n} \mathbb{1}_{\{\epsilon_l < \epsilon_i\}} - \mathbb{1}_{\{\hat{\epsilon}_l < \hat{\epsilon}_i\}} \right|. \tag{4.11}$$

We show the according box plots on the right side of Figure 4.6. As desired, this quantity approaches 0 for increasing sample size. For simplicity, we calculate this quantity only on a single split per simulation run.

4.7 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 dataset. 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 4.2, 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.

Acknowledgement

Christoph Schultheiss thanks Mathieu Chevalley for helpful discussions regarding finding use cases in the K562 dataset.

4.A Proofs

4.A.1 Proof of Theorem 4.1

Recall

$$Y \coloneqq \mathbb{E}[Y|\mathbf{X}] + \mathcal{E}.$$

Due to (A4.2), we have

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

Using (A4.1), $\mathbf{X}_U \perp \mathbf{H}_{\mathrm{PA}(Y)}|\mathbf{X}_{-U}$, and trivially, $\mathbf{X}_U \perp \mathbf{X}_{-U}|\mathbf{X}_{-U}$. It follows

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

Consider the counterfactual intervention. As $\mathbf{X}_{-U}$ remains unchanged, the second summand in (A4.2) 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 (A4.1), 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 (B4.1), 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.

4.A.2 Proof of Theorem 4.2

Consider first the $\subseteq$-statement. This means that $H_{0,j}$ in (4.6) must hold $\forall j \in W$. Let $S = \{1, \dots, p\} \setminus W$. 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.

4.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

$$\begin{aligned}
Q(\mathcal{E}, \mathbf{X}_U) &= \int \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t | \mathbf{X}_U)) d\mu(t) \geq 0 \\
S(\mathcal{E}) &= \int \mathrm{Var}(\mathbb{1}_{\mathcal{E} \geq t}) d\mu(t) \geq 0 \\
T(\mathcal{E}, \mathbf{X}_U) &= Q(\mathcal{E}, \mathbf{X}_U) / S(\mathcal{E}) \in [0, 1] \\
Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) &= \int \mathbb{E}[\mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t | \mathbf{X}_U, \mathbf{X}_S) | \mathbf{X}_S)] d\mu(t) \geq 0 \\
S(\mathcal{E}, \mathbf{X}_S) &= \int \mathbb{E}[\mathrm{Var}(\mathbb{1}_{\mathcal{E} \geq t} | \mathbf{X}_S)] d\mu(t) \geq 0 \\
T(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) &= Q(\mathcal{E}, \mathbf{X}_U | \mathbf{X}_S) / S(\mathcal{E}, \mathbf{X}_S) \in [0, 1].
\end{aligned}$$

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

$$\begin{aligned}
Q_n(\boldsymbol{\epsilon}, \mathbf{x}_U) &= \frac{1}{n^2} \sum_{i=1}^{n} \min\{R_i, R_{M(i)}\} - \frac{L_i^2}{n} \\
S_n(\boldsymbol{\epsilon}) &= \frac{1}{n^3} \sum_{i=1}^{n} L_i (n - L_i) \\
T_n(\boldsymbol{\epsilon}, \mathbf{x}_U) &= Q_n(\boldsymbol{\epsilon}, \mathbf{x}_U) / S_n(\boldsymbol{\epsilon})
\end{aligned}$$

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

4.A.4 Proof of Proposition 4.1

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 $d\nu(t) = 2d\mu(t)\ \forall t \geq 0$.

$$\begin{aligned}
Q(|\mathcal{E}|, \mathbf{X}_S) &= \int_0^\infty \mathrm{Var}(\mathbb{P}(|\mathcal{E}| \geq t|\mathbf{X}_S))d\nu(t) = \int_0^\infty \mathrm{Var}(2\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))2d\mu(t) \\
&= 8\int_0^\infty \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t), \\
&\quad \int_{-\infty}^0 \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t) = \int_{-\infty}^0 \mathrm{Var}(\mathbb{P}(\mathcal{E} \leq -t|\mathbf{X}_S))d\mu(t) \\
&= \int_{-\infty}^0 \mathrm{Var}(1 - \mathbb{P}(\mathcal{E} \geq -t|\mathbf{X}_S))d\mu(t) = \int_{-\infty}^0 \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq -t|\mathbf{X}_S))d\mu(t) \\
&\overset{t' \leftarrow -t}{=} \int_0^\infty \mathrm{Var}\big(\mathbb{P}\big(\mathcal{E} \geq t'|\mathbf{X}_S\big)\big)d\mu\big(t'\big) \text{ such that} \\
Q(\mathcal{E}, \mathbf{X}_S) &= \int_{-\infty}^\infty \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t) = 2\int_0^\infty \mathrm{Var}(\mathbb{P}(\mathcal{E} \geq t|\mathbf{X}_S))d\mu(t).
\end{aligned}$$

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.

4.A.5 Proof of Theorem 4.3

We build up the proof by some supporting Lemmata.

Lemma 4.1. Assume (B4.3) and (B4.4).

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

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 4.2. Assume the conditions of Lemma 4.1. 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

$$\mathbb{P}(|Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U) - \mathbb{E}[Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)]| \geq t) \leq K_1 \exp\big(-K_2 n t^2\big), \tag{4.12}$$

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

Lemma 4.3. 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 (B4.3) - (B4.4). 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 (B4.2) and (B4.5) 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 4.3 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 4.3, 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.

4.A.5.1 Proof of Lemma 4.1

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.

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

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.,

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

4.A.5.2 Proof of Lemma 4.2

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 $\hat{\boldsymbol{\epsilon}}$. 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 $\hat{R}_{M(i)}$.

$$
\begin{aligned}
&|Q_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - Q_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)| = \left| \frac{1}{n^2} \sum_{i=1}^{n} \min\{R_i, R_{M(i)}\} - \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} + \frac{\hat{L}_i^2 - L_i^2}{n} \right| \leq \\
&\left| \frac{1}{n^2} \sum_{i=1}^{n} \min\{R_i, R_{M(i)}\} - \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} \right| + \left| \frac{1}{n^3} \sum_{i=1}^{n} \hat{L}_i^2 - L_i^2 \right|. \\
&|S_n(g(\boldsymbol{\epsilon}), \mathbf{x}_U) - S_n(g(\hat{\boldsymbol{\epsilon}}), \mathbf{x}_U)| = \left| \frac{1}{n^2} \sum_{i=1}^{n} R_i - \hat{R}_i + \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} - \min\{R_i, R_{M(i)}\} \right| \leq \\
&\left| \frac{1}{n^2} \sum_{i=1}^{n} R_i - \hat{R}_i \right| + \left| \frac{1}{n^2} \sum_{i=1}^{n} \min\left\{\hat{R}_i, \hat{R}_{M(i)}\right\} - \min\{R_i, R_{M(i)}\} \right|. \\
&|S_n(g(\boldsymbol{\epsilon})) - S_n(g(\hat{\boldsymbol{\epsilon}}))| = \left| \frac{1}{n^3} \sum_{i=1}^{n} n\left(L_i - \hat{L}_i\right) + \hat{L}_i^2 - L_i^2 \right| \leq \left| \frac{1}{n^2} \sum_{i=1}^{n} L_i - \hat{L}_i \right| + \left| \frac{1}{n^3} \sum_{i=1}^{n} \hat{L}_i^2 - L_i^2 \right|.
\end{aligned}
$$

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.

$$\mathbb{E}\left[\left|\frac{1}{n^2}\sum_{i=1}^{n} R_i - \hat{R}_i\right|\right] \leq \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left|R_i - \hat{R}_i\right|\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left|\sum_{l=1}^{n} \mathbb{1}_{\{g(\epsilon_l)\leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l)\leq g(\hat{\epsilon}_i)\}}\right|\right] =$$
$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left|\sum_{l\neq i} \mathbb{1}_{\{g(\epsilon_l)\leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l)\leq g(\hat{\epsilon}_i)\}}\right|\right] \leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l\neq i}\mathbb{E}\left[\left|\mathbb{1}_{\{g(\epsilon_l)\leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l)\leq g(\hat{\epsilon}_i)\}}\right|\right] =$$
$$\frac{n^2-n}{n^2}\mathbb{E}\left[\left|\mathbb{1}_{\{g(\epsilon_l)\leq g(\epsilon_i)\}} - \mathbb{1}_{\{g(\hat{\epsilon}_l)\leq g(\hat{\epsilon}_i)\}}\right|\right] =$$
$$\frac{n^2-n}{n^2}\mathbb{P}([g(\epsilon_l)\leq g(\epsilon_i)\cap g(\hat{\epsilon}_l) > g(\hat{\epsilon}_i)]\cup[g(\epsilon_l)\geq g(\epsilon_i)\cap g(\hat{\epsilon}_l) < g(\hat{\epsilon}_i)]) \overset{n\to\infty}{\to} 0$$

by Lemma 4.1. 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 4.1, the bound on the convergence rate follows directly.

4.A.5.3 Proof of Lemma 4.3

By Lemma 4.2, there exists a rate, say, $D(n) = \mathcal{O}(1)$ such that

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

Then,

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

using the rate derived in Lemma 14.2 of (Azadkia and Chatterjee, 2021). Therefore, with (4.12),

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

4.A.6 Proof of Theorem 4.4

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

Using Lemma 4.2 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

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

where we used Lemma 14.2 in (Azadkia and Chatterjee, 2021) in the second to last inequality. Fi-

nally, 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.

4.A.7 Proof of Proposition 4.2

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,

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

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

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

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

We consider the only random term

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n \min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\}\right] &= \mathbb{E}\left[\sum_{\hat{R}_i=1}^n \min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\}\right] = \mathbb{E}\left[\sum_{\hat{R}_i=1}^n \min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\} | \hat{R}_1, \dots, \hat{R}_n\right] \\ &= \sum_{\hat{R}_i=1}^n \mathbb{E}\Big[\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\} | \hat{R}_1, \dots, \hat{R}_n\Big] = \sum_{\hat{R}_i=1}^n \mathbb{E}\Big[\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\} | \hat{R}_i\Big].\end{aligned}$$

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 $\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\}$. We analyse the expectation under the approximate conditional distribution as given above. If $\hat{R}_i \leq n/k$, $\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\} = \hat{R}_i$. If $n/k < \hat{R}_i \leq 2n/k$ and

the uniformly chosen number is 0, $\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\} = \hat{R}_i - n/k$. This has probability $1/k$. Otherwise, $\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\} = \hat{R}_i$. Analogously, if $ln/k < \hat{R}_i \leq (l+1)n/k$ for some integer $0 \leq l < k$, i.e., $l = \max\Big\{r \in \mathbb{N}_0 | r < \hat{R}_i k/n\Big\}$, it holds under the approximate distribution

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

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

$$\sum_{\hat{R}_i=1}^{n}\tilde{\mathbb{E}}\Big[\min\Big\{\hat{R}_i, \hat{R}_{M(i)}\Big\}|\hat{R}_i\Big] = \sum_{\hat{R}_i=1}^{n}\hat{R}_i - \frac{n^2}{2k^3}\sum_{l=0}^{k-1}\big(l^2+l\big) =$$
$$\frac{n^2+n}{2} - \frac{n^2}{2k^3}\left(\frac{k(k-1)(2k-1)}{6} + \frac{k(k-1)}{2}\right) = \frac{n^2+n}{2} - \frac{n^2}{6k^3}\big(k^3-k\big).$$

Then,

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

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

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

We even control

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

For arbitrary conditioning events A, we have

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

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

Hence, we can ignore events with vanishing probability. Let $v_1,\dots,v_k$ 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

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

Under A^c, there are only $\mathcal{O}\big(n^{3/4}\big)$ ranks $\hat{R}_i$ for which $\epsilon_i\neq v_{l+1}$ with $l=\max\Big\{r\in\mathbb{N}_0|r<\hat{R}_ik/n\Big\}$ is possible. Summing over these leads to another $\mathcal{o}\big(n^2\big)$ term and can be ignored. Consider the $\hat{R}_i$ for which $A_i\coloneqq\{\epsilon_i=v_{l+1}\}$ 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 $n_1,\dots,n_k$, we have

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

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

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

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

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

$$\begin{aligned}\min\Big\{\hat{R}_i,\hat{R}_{M(i)}\Big\}=\hat{R}_{M(i)}&=\sum_{r=1}^{m}n_r+\sum_{l:\epsilon_l=v_{m+1}}\mathbb{1}_{\{x_l\leq x_{M(i)}\}}\\&=\sum_{r=1}^{m}n_r+\sum_{l:\epsilon_l=v_{m+1}}\mathbb{1}_{\{x_l\leq x_i\}}+\mathbb{1}_{\{x_{M(i)}>x_i\}}.\end{aligned}$$

Under the given conditioning, this is

$$\begin{aligned}&\sum_{r=1}^{m}n_r+n_{m+1}\left(\frac{\hat{R}_i-\sum_{r=1}^{l}n_r}{n_{l+1}}+\mathcal{O}(1)\right)+\mathcal{O}(1)=\\&\frac{nm}{k}+\mathcal{O}\Big(n^{3/4}\Big)+\Big(\frac{n}{k}+\mathcal{O}\Big(n^{3/4}\Big)\Big)\frac{\hat{R}_i-nl/k+\mathcal{O}\big(n^{3/4}\big)}{n/k+\mathcal{O}\big(n^{3/4}\big)}+\mathcal{O}(n)=\\&\frac{nm}{k}+\Big(1+\mathcal{O}\Big(n^{-1/4}\Big)\Big)\left(\hat{R}_i-\frac{nl}{k}+\mathcal{O}\Big(n^{3/4}\Big)\right)+\mathcal{O}(n)=\frac{nm}{k}+\hat{R}_i-\frac{nl}{k}+\mathcal{O}\Big(n^{3/4}\Big)+\mathcal{O}(n)=\\&\hat{R}_i-\frac{n(l-m)}{k}+\mathcal{O}(n).\end{aligned}$$

In summary,

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

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

4.A.8 Proof of Theorem 4.5

Due to (A4.2*), we have

$$\begin{aligned}\mathbb{E}[Y|\mathbf{X}]=&f_{\mathbf{X}_UY}\big(\mathbf{X}_U,\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)+g_{\mathbf{X}_UY}\big(\mathbf{X}_U,\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)\mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big],\\\mathrm{Var}(Y|\mathbf{X})=&g^2_{\mathbf{X}_UY}\big(\mathbf{X}_U,\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)\\&\Big(\mathbb{E}\big[f^2_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big]-\mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big]^2\Big),\\\mathcal{E}=&\frac{f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)-\mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big]}{\sqrt{\mathbb{E}\big[f^2_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big]-\mathbb{E}\big[f_{\mathbf{H}Y}\big(\mathbf{H}_{\mathrm{PA}(Y)},\mathbf{X}_{\mathrm{PA}(Y)\setminus U}\big)|\mathbf{X}\big]^2}}.\end{aligned}$$

As in Section 4.A.1,

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

(A4.1) together with (B4.1) implies that only terms involving $\mathbf{X}_U$ can be different for the counterfactual; see Section 4.A.1. In particular, $\mathcal{E}$ cannot change. Hence, Y changes from

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

which is as stated in the theorem.

4.A.9 Proof of Theorem 4.6

We have the following supporting result.

Lemma 4.4. Suppose that (B4.6) - (B4.8) hold. Then

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

With Lemma 4.4 we have replaced Assumption (B4.3) which is the only missing part to reconstruct the asymptotic results as in Theorems 4.3 and 4.4.

4.A.9.1 Proof of Lemma 4.4

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

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

Consider the difference

$$\begin{aligned}\Big|V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)\Big| &= \Big|f_2(\mathbf{x}_i) - (f_2(\mathbf{x}_i) + \mathcal{O}_p(1)) - f_1^2(\mathbf{x}_i) + (f_1(\mathbf{x}_i) + \mathcal{O}_p(1))^2\Big| \\ &\leq \mathcal{O}_p(1) + |f_1(\mathbf{x}_i)|\mathcal{O}_p(1) = \mathcal{O}_p(1).\end{aligned}$$

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,

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

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 $\eta > 0$.

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0\right) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 | \hat{V}(\mathbf{x}_i) > 0\right)\mathbb{P}\Big(\hat{V}(\mathbf{x}_i) > 0\Big) + \\
&\qquad \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 | \hat{V}(\mathbf{x}_i) \leq 0\right)\mathbb{P}\Big(\hat{V}(\mathbf{x}_i) \leq 0\Big) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta \cup \hat{V}(\mathbf{x}_i) < 0 | \hat{V}(\mathbf{x}_i) > 0\right) + \mathbb{P}\Big(\hat{V}(\mathbf{x}_i) \leq 0\Big) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\left|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\right| \geq \eta | \hat{V}(\mathbf{x}_i) > 0\right) = \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)} + \sqrt{\hat{V}(\mathbf{x}_i)}}\right| \geq \eta | \hat{V}(\mathbf{x}_i) > 0\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)}}\right| \geq \eta | \hat{V}(\mathbf{x}_i) > 0\right) \leq \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{V(\mathbf{x}_i) - \hat{V}(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)}}\right| \geq \eta\right) / \mathbb{P}\Big(\hat{V}(\mathbf{x}_i) > 0\Big) = 0.
\end{aligned}$$

Consider the residuals assuming a positive variance estimate.

$$\begin{aligned}
|\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i| &= \left|\frac{y_i - \hat{f}_i(\mathbf{x}_i)}{\sqrt{\hat{V}(\mathbf{x}_i)}} - \frac{y_i - f_i(\mathbf{x}_i)}{\sqrt{V(\mathbf{x}_i)}}\right| = \left|\frac{\Big(y_i - \hat{f}_i(\mathbf{x}_i)\Big)\sqrt{V(\mathbf{x}_i)} - (y_i - f_i(\mathbf{x}_i))\sqrt{\hat{V}(\mathbf{x}_i)}}{\sqrt{\hat{V}(\mathbf{x}_i)}\sqrt{V(\mathbf{x}_i)}}\right| \\
&= \left|\frac{\Big(f_i(\mathbf{x}_i) - \hat{f}_i(\mathbf{x}_i)\Big)\sqrt{V(\mathbf{x}_i)} + (y_i - f_i(\mathbf{x}_i))\Big(\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\Big)}{\sqrt{\hat{V}(\mathbf{x}_i)}\sqrt{V(\mathbf{x}_i)}}\right| \\
&\leq \frac{1}{\sqrt{\hat{V}(\mathbf{x}_i)}}\left(\Big|f_i(\mathbf{x}_i) - \hat{f}_i(\mathbf{x}_i)\Big| + |\epsilon_i|\Big|\sqrt{V(\mathbf{x}_i)} - \sqrt{\hat{V}(\mathbf{x}_i)}\Big|\right).
\end{aligned}$$

Arguing similarly as before, we have

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

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

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

Chapter 5

On the pitfalls of Gaussian likelihood scoring for causal discovery

Christoph Schultheiss and Peter Bühlmann
Journal of Causal Inference, 11(1):20220068.

Abstract

We consider likelihood score-based methods for causal discovery in structural causal models. In particular, we focus on Gaussian scoring and analyze the effect of model misspecification in terms of non-Gaussian error distribution. We present a surprising negative result for Gaussian likelihood scoring in combination with nonparametric regression methods.

5.1 Introduction

We consider the problem of finding the causal structure of a set of random variables $X_1, \dots X_p$. We assume that the data can be represented by a structural equation model whose structure is given by a directed acyclic graph (DAG), say, G^0, with nodes $1, \dots, p$ and denote by $\mathrm{PA}(j)$ the parents of j. Without further assumptions on the structural causal model, one can only find G^0 up to its Markov equivalence class. This can, e.g., be achieved with the PC algorithm (Spirtes et al., 2000). Its generality comes at the price of requiring conditional independence tests, which are a statistically hard problem.

Here, we focus on the so-called additive noise model (ANM)

$$X_j \leftarrow f_j\big(X_{\mathrm{PA}(j)}\big) + \mathcal{E}_j \quad \forall j \in 1, \dots, p, \tag{5.1}$$

where the $\mathcal{E}_j$ are mutually independent centered random variables. This is a popular modelling assumption that allows for better identifiability guarantees; see, e.g., (Hoyer et al., 2008a; Peters et al., 2014).

For an arbitrary DAG G, let $\mathrm{PA}^G(j)$ be the nodes from which a directed edge towards j starts. Define

$$\mathcal{E}_j^G = X_j - \mathbb{E}\Big[X_j | X_{\mathrm{PA}^G(j)}\Big] \quad \forall j \in 1, \dots, p.$$

Obviously, $\mathcal{E}_j^{G^0} = \mathcal{E}_j$. Under not overly restrictive assumptions, it holds that $\mathcal{E}_1^G, \dots \mathcal{E}_p^G$ are mutually independent only if $G \supseteq G^0$, see Peters et al. (2014). Thus, an obvious approach to find G^0 is to loop over all possible graphs and test for independence of the residuals. Of course, this becomes infeasible when the dimensionality p grows.

A more reasonable algorithm based on greedy search is presented in Peters et al. (2014). They also introduce RESIT (regression with subsequent independence test) that iteratively detects sink nodes. Finding the true DAG is guaranteed assuming perfect regressors and independence tests. It involves $O\big(p^2\big)$ nonparametric independence tests which might be computationally involved or lacking power when the sample size is small.

Instead of performing independence tests, one can compare the likelihood score of different graphs

$$\mathcal{L}(G) = \mathbb{E}\left[\log\left(\prod_{j=1}^{p} p_j^G\big(\mathcal{E}_j^G\big)\right)\right] = \sum_{j=1}^{p} \mathbb{E}\big[\log\big(p_j^G\big(\mathcal{E}_j^G\big)\big)\big],$$

where p_j^G denotes the density of $\mathcal{E}_j^G$. Only for independent $\mathcal{E}_j^G$, their multivariate density factorizes as suggested. Therefore, the true DAG maximizes this quantity due to the properties of the Kullback-Leibler divergence. In practice, this comes with the additional difficulty of estimating the densities

$p_j^G(\cdot)$ and one needs to add some penalization to prefer simpler graphs and avoid selecting complete graphs; see, e.g., Nowzohour and Bühlmann (2016).

If one additionally assumes that the $\mathcal{E}_j$ are marginally normally distributed $\mathcal{N}\Big(0, \sigma_j^2\Big)$ one can instead consider the Gaussian likelihood

$$\mathcal{L}^{\mathcal{N}}(G) \coloneqq \sum_{j=1}^{p}\Big(-\log\big(\sigma_j^G\big) - \frac{1}{2} - \frac{1}{2}\log(2\pi)\Big) = -\sum_{j=1}^{p}\log\big(\sigma_j^G\big) + C,$$

where $\Big(\sigma_j^G\Big)^2 = \mathbb{E}\Big[\Big(\mathcal{E}_j^G\Big)^2\Big]$ (Bühlmann et al., 2014). It holds $\mathcal{L}(G) \geq \mathcal{L}^{\mathcal{N}}(G)$ and $\mathcal{L}^{\mathcal{N}}\big(G^0\big) = \mathcal{L}\big(G^0\big)$. Thus, the problem simplifies to finding the graph that leads to the lowest sum of log-variances.

Under such a normality assumption for $\mathcal{E}_j$ and the $f_j(\cdot)$ in (5.1) being additive in their arguments, Bühlmann et al. (2014) present a causal discovery method that is consistent for high-dimensional data. To justify this approach for a broader class of error distributions, define

$$\Delta \coloneqq \min_{G \not\supseteq G^0} \sum_{j=1}^{p} \log\big(\sigma_j^G\big) - \log\Big(\sigma_j^{G^0}\Big) \tag{5.2}$$

Then, one has to assume that

(A5.1) $\Delta > 0$.

That is, the lowest possible expected negative Gaussian log-likelihood with any graph G that is not a superset of the true G^0 must be strictly larger than the expected negative Gaussian log-likelihood with the true graph G^0. An argument for that assumption is that a true causal model should be easier to fit in some sense and thus also obtain lower error variance. Using simple "non-pathological" examples, we demonstrate that this can easily be a fallacy when the true error distribution is misspecified. Thus, we advocate the need to be very careful when using Gaussian scoring with flexible nonparametric regression functions in causal discovery. The main part of this paper considers theoretical population properties. We discuss some data applications in Section 5.5.1.

5.2 Data-generating linear model

We consider first data-generating linear models where

$$f_j\big(X_{PA(j)}\big) = \sum_{k \in PA(j)} \beta_{jk} X_k.$$

For these, we find the explicit Theorem 5.1. The intuition for this result carries over to a range of nonlinear ANM (5.1), especially if the causal effects are close to linear. We present according examples in Section 5.3.

If all the $\mathcal{E}_j$ in a data-generating linear model are Gaussian, i.e., $X_1, \ldots, X_p$ are jointly Gaussian, it is known that any causal order could induce the given multivariate normal distribution and obtain an optimal Gaussian score. Assuming that the distribution is faithful with respect to the true DAG G^0, the set of the most sparse DAGs obtaining the optimal Gaussian score corresponds to the Markov equivalence class of G^0; see, e.g., Zhang and Spirtes (2008). Thus, one can obtain this Markov equivalence class by preferring more sparse DAGs in case of equal scores. In general, the single true DAG cannot be determined even if the full multivariate distribution is known. On the contrary, the Linear Non-Gaussian Acyclic Model (LiNGAM) introduced in Shimizu et al. (2006) is known to be identifiable. In such linear non-Gaussian models, algorithms designed for linear Gaussian models, e.g., the PC algorithm using partial correlation to assess conditional independence, do not use all the available information, but typically still provide the same guarantees since they depend on the covariance structure only. The covariance matrix of data generated by a linear causal model does not change when replacing a Gaussian $\mathcal{E}_j$ by an additive error of the same variance but otherwise arbitrary distribution. Thus, assuming the faithfulness condition for the data-generating distribution, Gaussian scoring for linear causal models in the infinite sample limit leads to the true underlying Markov equivalence class even under misspecification of the error distribution.

If the data-generating model is not known to be linear such that nonparametric regression methods, or the conditional mean as their population version, are applied, this generalization does not hold true anymore, as laid out in the following theorem. Let π be a permutation on $\{1, \ldots, p\}$ and G^π the full DAG according to π, i.e., $\pi(k) \in \mathrm{PA}^{G^\pi}(\pi(j))$ if and only if $k < j$.

Theorem 5.1. Let $X_1, \ldots, X_p$ come from a linear model:

$$X_j \leftarrow \sum_{k \in PA(j)} \beta_{jk} X_k + \mathcal{E}_j, \quad \text{with} \quad \mathbb{E}[\mathcal{E}_j] = 0, \quad \mathbb{E}\big[\mathcal{E}_j^2\big] < \infty \quad \forall j \in 1, \ldots, p, \tag{5.3}$$

with mutually independent $\mathcal{E}_j$. Then, for every possible permutation π,

$$\sum_{j=1}^{p} \log\big(\sigma_j^{G^\pi}\big) - \log\Big(\sigma_j^{G^0}\Big) \leq 0.$$

That is, for every causal order, the corresponding full graph scores at least as well as the true causal graph.
Furthermore,

$$\sum_{j=1}^{p} \log\big(\sigma_j^{G^\pi}\big) - \log\Big(\sigma_j^{G^0}\Big) < 0 \quad \text{if}$$
$$\exists j \in \{1, \ldots, p\} : \quad \mathbb{E}\Big[X_j | X_{PA^{G^\pi}(j)}\Big] \neq X_{PA^{G^\pi}(j)}^\top \beta^{\pi, j} \quad \text{where}$$

$$\beta^{\pi,j} = \mathbb{E}\Big[X_{PA^{G^\pi}(j)} X_{PA^{G^\pi}(j)}^\top\Big]^{-1} \mathbb{E}\Big[X_j X_{PA^{G^\pi}(j)}\Big] \quad \text{is the least squares parameter.}$$

That is, for every causal order, the corresponding full graph scores strictly better than the true causal graph if at least one conditional expectation is nonlinear in the parental variables.

Apart from some pathological cases, the last condition holds for permutations that are not conformable with the true DAG unless all the $\mathcal{E}_j$ are Gaussian. Thus, the gap condition (A5.1) does not hold true for this whole set of distributions, and, without Gaussianity, a wrong model would not only score equivalently but even be preferred. This is in stark contrast to results around Gaussian scoring when fitting linear models.

5.2.1 Illustrative examples

For illustrative purposes, we restrict ourselves to the two variable case with $X_2 \leftarrow \beta X_1 + \mathcal{E}_2$. We have the true DAG $G^0 = \{X_1 \to X_2\}$ and define $G^\perp = \{X_1 \leftarrow X_2\}$. If $\beta X_1 \overset{\mathcal{D}}{=} \mathcal{E}_2$, it holds

$$X_2 = \mathbb{E}[X_2|X_2] = \mathbb{E}[\beta X_1 + \mathcal{E}_2|X_2] = 2 * \beta \mathbb{E}[X_1|X_2] \quad \text{implying that} \quad \mathbb{E}[X_1|X_2] = \frac{1}{2\beta} X_2$$

such that $\Delta = 0$; see the definition in (5.2). In the two variable linear model, $\exp(\Delta)^2$ equals the ratio of the attainable mean squared error for the backward direction between the best-fitting unrestricted model and the best-fitting linear model.

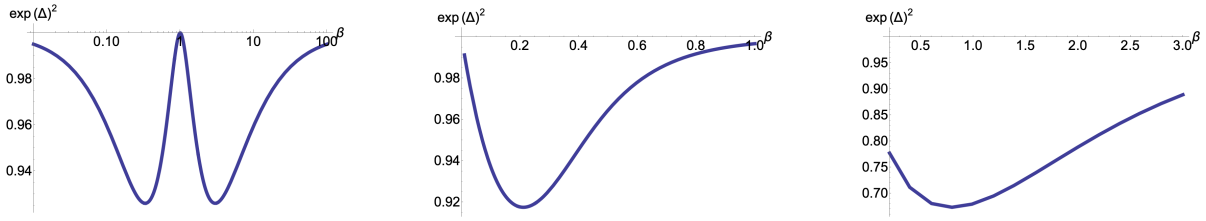
Consider first the analytically tractable case where $X_1 \overset{\mathcal{D}}{=} \mathcal{E}_2 \sim \text{Unif}[-1, 1]$. Then, for $\beta = \pm 1$, $\mathbb{E}[X_1|X_2] = \pm X_2/2$ and $\Delta = 0$. For every other nonzero and bounded β, $\Delta < 0$ so a causal discovery method based on Gaussian scoring would - assuming a correct regression function estimator - wrongly claim $X_2 \to X_1$ in the large sample limit. We make things more explicit.

Proposition 5.1. Let $X_2 = \beta X_1 + \mathcal{E}_2$ with $X_1 \overset{\mathcal{D}}{=} \mathcal{E}_2 \sim \text{Unif}[-1, 1]$. Define $\gamma = \max\{|\beta|, |1/\beta|\}$. Then, the ratio of the variances is given by

$$\exp(\Delta)^2 = \left(\prod_{j=1}^{2} \left(\sigma_j^{G^\perp}\right)^2\right) / \left(\prod_{j=1}^{2} \left(\sigma_j^{G^0}\right)^2\right) = \frac{(\gamma^2 + 1)(2\gamma - 1)}{2\gamma^3} := r(\gamma).$$

As argued above, $r(1) = 1$. For $|\beta| \to 0$, the ratio between the variance products approaches 1 as X_1 and X_2 become independent, and, hence, both models perform equally well. For $|\beta| \to \infty$, the ratio approaches 1 as X_2 becomes a deterministic linear map of X_1, and, hence, the linear model is invertible. The ratio $r(\gamma)$ is minimized for $\gamma = 3$, with $r(3) \approx 0.93$, strictly decreasing for $\gamma \in [1, 3)$, and strictly increasing for $\gamma \in (3, \infty)$. This is visualized in Figure 5.1a.

Consider next a similar example but with $X_1 \sim \mathcal{N}(0, 1)$. Analytic expressions for $\mathbb{E}[X_1|X_2]$ and $\text{Var}(X_1|X_2)$ can be found in terms of the Gaussian cumulative distribution function. For



(a) $X_1 \stackrel{\mathcal{D}}{=} \mathcal{E}_2 \sim \text{Unif}[-1, 1]$ (b) $X_1 \sim \mathcal{N}(0, 1)$, $\mathcal{E}_2 \sim \text{Unif}[-1, 1]$ (c) $X_1 \sim \mathcal{N}(0, 1)$, $\mathcal{E}_2 + 1 \sim \chi^2_1$

Figure 5.1: Two variable linear model: effect of changing β.

$\left(\sigma_1^{G^\perp}\right)^2 = \mathbb{E}[\text{Var}(X_1|X_2)]$ we invoke numerical integration. The obtained $\exp(\Delta)^2$ is shown in Figure 5.1b. As elaborated before, it approaches 1 for $\beta \to 0$ and $\beta \to \infty$. In between, it is strictly less than 1 since $\mathbb{E}[X_1|X_2]$ is not linear in X_2. The minimum of $\exp(\Delta)^2 \approx 0.92$ is obtained for $\beta \approx 0.21$.

Finally, we use an asymmetric error distribution instead, namely, $\mathcal{E}_2 \sim \chi^2_1 - 1$. Figure 5.1c shows that this leads to more extreme values of $\Delta < 0$. Notably, $\mathbb{E}[X_1|X_2]$ is not monotone in this case so the best linear fit is not a good approximation.

5.3 Beyond a data-generating linear model

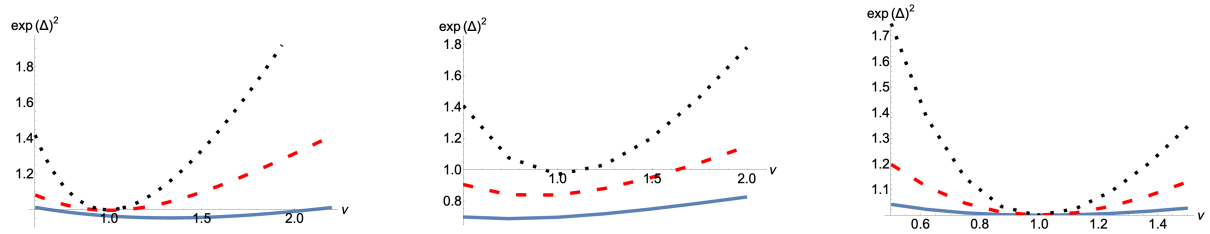
If all $\mathcal{E}_j$ are Gaussian, we know that $\Delta = 0$ for linear conditional expectations in (5.1), but $\Delta > 0$ otherwise. Thus, involving nonlinearities enables the identifiability of the model.

For non-Gaussian $\mathcal{E}_j$ in a linear model, $\Delta < 0$ holds true apart from some special cases; see, Theorem 5.1. The intuition is that nonlinearities could be beneficial for the identifiability nevertheless. As the lower bound for Δ is negative and not 0, presumably a higher degree of nonlinearity might be necessary to achieve $\Delta > 0$. We analyze this with the following simple model

$$X_2 \leftarrow \beta \text{Sign}(X_1) \frac{|X_1|^\nu}{\sqrt{\mathbb{E}\left[|X_1|^{2\nu}\right]}} + \mathcal{E}_2, \quad \nu > 0. \tag{5.4}$$

The normalization ensures that the variance of X_2 depends only on β. For $\nu = 1$, we obtain a linear model.

Consider the case $X_1 \sim \mathcal{N}(0, 1)$ and $\mathcal{E}_2 \sim \text{Unif}[-1, 1]$. Analytic expressions for $\mathbb{E}[X_1|X_2]$ and $\text{Var}(X_1|X_2)$ can be found in terms of the Gaussian cumulative distribution function and the Γ-function. For $\left(\sigma_1^{G^\perp}\right)^2 = \mathbb{E}[\text{Var}(X_1|X_2)]$ we invoke numerical integration. The obtained $\exp(\Delta)^2$ is shown in Figure 5.2a for different values of β. It confirms the intuition, that sufficiently strong nonlinearity leads to $\Delta > 0$ even for non-Gaussian errors. For $\beta = 1$ and $\beta = 2$, the model becomes identifiable for most $\nu \neq 1$. For $\beta = 0.5$, stronger nonlinearities are necessary. Also, the minimum Δ is not obtained for $\nu = 1$ but at $\nu \approx 1.32$. Thus, not every nonlinear model is better identifiable than the corresponding linear model.



(a) $X_1 \sim \mathcal{N}(0,1)$, $\mathcal{E}_2 \sim \text{Unif}[-1,1]$ (b) $X_1 \sim \mathcal{N}(0,1)$, $\sqrt{6}\mathcal{E}_2 + 1 \sim \chi^2_1$ (c) $X_1 \overset{\mathcal{D}}{=} \sqrt{3}\mathcal{E}_2 \sim \mathcal{N}(0,1)$

Figure 5.2: Two variable nonlinear model (5.4): effect of changing ν for $\beta = 0.5$ (solid blue curve), $\beta = 1$ (dashed red curve), and $\beta = 2$ (dotted black curve).

As in the linear case, we consider the effect of an assymetric error distribution, namely, a scaled and centered chi-squared distribution. We show this in Figure 5.2b. The factor $\sqrt{6}$ leads to the different lines corresponding to the same respective signal-to-noise ratios. As expected, higher degrees of nonlinearity are necessary to obtain a positive Δ.

We show the behavior for a correctly specified model in 5.2c. For $\nu = 1$, $\Delta = 0$ as implied by the unidentifiability result, otherwise, $\Delta > 0$. For $\nu \neq 1$, higher signal-to-noise ratios lead to more distinct $\Delta > 0$.

Monotonicity of $f_2(\cdot)$ The nonlinearities discussed here are designed to be slight deviations from the linear model and, thus, chosen to be strictly monotone. Notably, for non-monotone functions, the intuition that the anti-causal model is harder to fit is more applicable. In particular, if X_1 is centered and symmetric, and $f_2(\cdot)$ is an even function, it holds $\mathbb{E}[X_1|X_2] \equiv 0$. Then,

$$\left(\sigma_1^{G^\perp}\right)^2 = \text{Var}(X_1) \quad \text{and} \quad \exp(\Delta)^2 = \frac{\text{Var}(X_1)\text{Var}(X_2)}{\text{Var}(X_1)\text{Var}(\mathcal{E}_2)} > 1. \tag{5.5}$$

Thus, the gap condition (A5.1) is satisfied regardless of the distribution of $\mathcal{E}_2$ as long as X_1 and X_2 have finite variance.

5.4 Heteroskedastic noise model

A simple extension of model (5.1), that has recently gained some attention, is the heteroskedastic noise model also referred to as location-scale noise model

$$X_j \leftarrow f_j\big(X_{\text{PA}(j)}\big) + g_j\big(X_{\text{PA}(j)}\big)\mathcal{E}_j \quad \text{with} \quad \mathbb{E}[\mathcal{E}_j] = 0, \quad \mathbb{E}\big[\mathcal{E}_j^2\big] = 1 \quad \forall j \in 1, \dots, p,$$

with some nonnegative functions $g_j(\cdot)$ (Strobl and Lasko, 2023; Xu et al., 2022; Immer et al., 2023), see also Chapter 6. It comes with similar identifiability guarantees as the ANM when testing for

mutual independence between the $\mathcal{E}_j$. Let accordingly

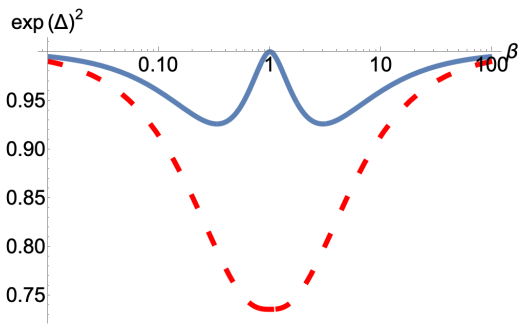
$$f_j^G\left(X_{\mathrm{PA}^G(j)}\right) = \mathbb{E}\left[X_j | X_{\mathrm{PA}^G(j)}\right], \quad g_j^G\left(X_{\mathrm{PA}^G(j)}\right)^2 = \mathbb{E}\left[\left(X_j - f_j^G\left(X_{\mathrm{PA}^G(j)}\right)\right)^2 | X_{\mathrm{PA}^G(j)}\right] \quad \text{and}$$

$$\mathcal{E}_j^G = \frac{X_j - f_j^G\left(X_{\mathrm{PA}^G(j)}\right)}{g_j^G\left(X_{\mathrm{PA}^G(j)}\right)}$$

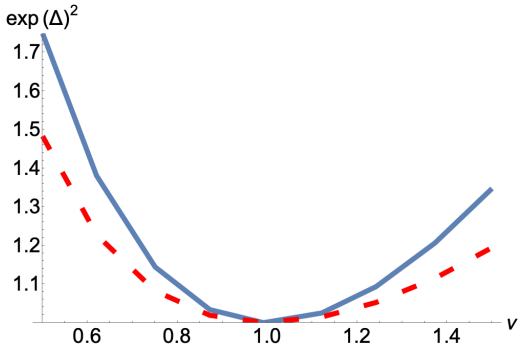
be the conditional means, conditional variances, and residuals according to any, potentially wrong, DAG G. Then, one gets

$$\begin{aligned}
\mathcal{L}(G) &= \sum_{j=1}^{p} \mathbb{E}\left[\log\left(\frac{p_j^G\left(\mathcal{E}_j^G\right)}{g_j^G\left(X_{\mathrm{PA}^G(j)}\right)}\right)\right] \\
\mathcal{L}^{\mathcal{N}}(G) &:= -\sum_{j=1}^{p} \mathbb{E}\left[\log\left(g_j^G\left(X_{\mathrm{PA}^G(j)}\right)\right)\right] + C = -\sum_{j=1}^{p} \frac{1}{2}\mathbb{E}\left[\log\left(g_j^G\left(X_{\mathrm{PA}^G(j)}\right)^2\right)\right] + C \\
&\geq -\sum_{j=1}^{p} \frac{1}{2}\log\left(\mathbb{E}\left[g_j^G\left(X_{\mathrm{PA}^G(j)}\right)^2\right]\right) + C = -\sum_{j=1}^{p} \log\left(\sigma_j^G\right) + C.
\end{aligned}$$

Thus, when fitting heteroskedastic models the score can only be increased compared to the homoskedastic fit. This can further increase the difficulty of finding the correct direction under non-Gaussian noise. Even if the true forward model is homoskedastic, i.e., $g_j(\cdot) \equiv \sigma_j$, the backward model is typically heteroskedastic and profits from this new score. For example, in the set-up of Figure 5.1a, Δ would be negative even for $\beta = 1$. If one allows to fit heteroskedastic models, a result analagous to Theorem 5.1 exists. A negative gap is obtained unless all conditional expectations are linear and all conditional variances are constant for the wrong causal order.



(a) Model (5.3) with $X_1 \overset{\mathcal{D}}{=} \mathcal{E}_2 \sim \mathrm{Unif}[-1, 1]$



(b) Model (5.4) with $X_1 \overset{\mathcal{D}}{=} \sqrt{3}\mathcal{E}_2 \sim \mathcal{N}(0, 1)$ and $\beta = 2$

Figure 5.3: Two variable additive model: fitting homoskedastic models (solid blue curve) versus heteroskedastic models (dashed red curve).

In Figure 5.3, we review the examples from Figures 5.1a and 5.2c and see how allowing for a heteroskedastic fit makes the problem harder. For the sake of comparison, we look at $\exp(\Delta)^2$ although it does not have the same simple interpretation in the heteroskedastic case.

In terms of the location-scale noise model, the data-generating model as in Figure 5.3a is unidentifiable as $X_1|X_2$ is uniformly distributed, i.e., its distribution is independent of X_2 apart from location and scale. This does not contradict the identifiability results as they are derived for random variables with full support in $\mathbb{R}$.

5.5 Discussion

5.5.1 Data applications

For an extensive comparison between methods relying on Gaussian scoring and nonparametric independence tests in additive noise models or heteroskedastic noise models, we refer to Immer et al. (2023). There, several fitting methods are considered and combined with both approaches and evaluated on a variety of benchmark cause and effect pairs. Those pairs include both real and artificial data. For some of the considered data sources, using independence tests clearly improved the success rate for inferring the causal direction as compared to using the Gaussian score.

Let us consider two specific examples of the Tübingen data by Mooij et al. (2016). Details on the data can be found in Section D.11 of their paper. Both examples have the temperature as the effect variable while the cause is the day of the year or the intensity of the solar radiation, respectively. We

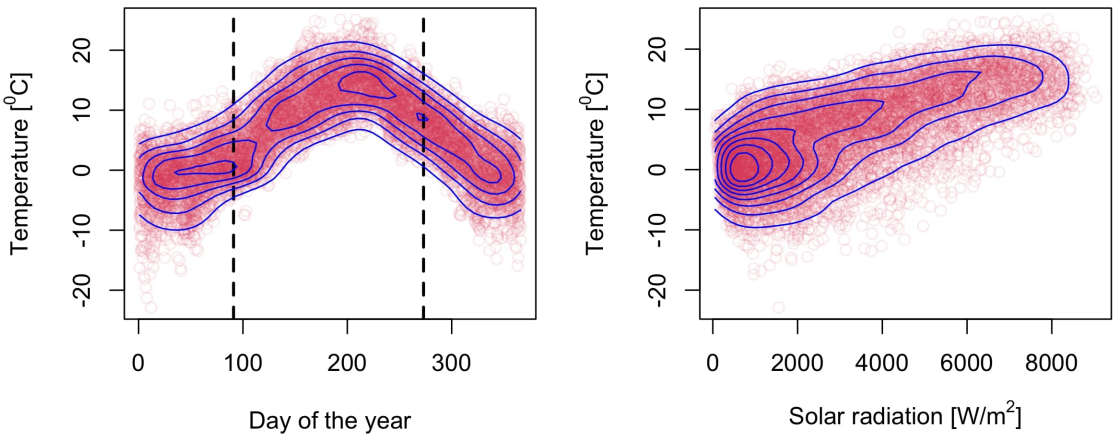


Figure 5.4: Scatter plot and contour lines of the density estimate for two selected pairs from the Tübingen data.

show the corresponding scatter plots as well as the contour lines of the density estimates in Figure 5.4. It is evident that in neither case the cause variable is normally distributed, the days are perfectly uniformly distributed while solar radiation is right-skewed. Therefore, the assumptions for Gaussian scoring to infer the true causal direction are not fulfilled. For the first data set, we restrict the numerical analysis to the time frame April 1st to September 30th (March 31st to September 29th in leap years) to circumvent the issue that the data are circular. This is indicated by the black dotted lines.

To evaluate the Gaussian scores, we estimate the conditional expectation for either direction with a smoothing spline. In the first case, the causal effect is non-monotone, and the conditional mean in the anti-causal direction is not very informative to predict the day of the year. Therefore, we obtain the correct causal direction with Gaussian scoring even though the assumptions are not fulfilled. We get the data estimate $\exp(\Delta)^2 = 1.48$.

The effect of solar radiation on the temperature appears to be monotone which makes the conditional expectation in the anti-causal direction more informative. Also, it seems that the conditional expectation in the causal direction is not so far from being linear. This indeed makes the Gaussian scoring algorithm prefer the wrong direction. The estimate is $\exp(\Delta)^2 = 0.91$. Similarly, it fails for the first data set when considering the first 183 days of the year instead ($\exp(\Delta)^2 = 0.99$).

With RESIT relying on independence testing, we see for both data sets that the hypothesis of residuals being independent of the predictor is rejected in either direction. This indicates that the ANM in (5.1) is not rich enough to explain the data. However, applying Algorithm 1 from Peters et al. (2014) which minimizes the estimated dependence between predictor and residuals finds the true causal direction for both data pairs.

5.5.2 Conclusion

We discuss causal discovery in structural causal models using Gaussian likelihood scoring and analyze the effect of model misspecification.

In the case where the data-generating distribution comes from a linear structural equation model and linear regression functions are used for estimation, the following holds. When the true error distribution is Gaussian, one can only identify the Markov equivalence class of the underlying data-generating DAG. The same is true when the error distribution is non-Gaussian but one wrongly relies on a Gaussian error distribution for estimation.

Thus, popular algorithms like the greedy equivalence search (GES) (Chickering, 2002) for Gaussian models or the PC algorithm (Spirtes et al., 2000) assessing partial correlation are potentially conservative and only infer the Markov equivalence class when the error distributions are non-Gaussian as they do not exploit the maximal amount of information. But they are safe to use within the domain of data-generating linear structural equation models. We prove here that this fact does not necessarily

hold true when invoking nonparametric regression estimation. Especially, if the true causal model is linear or just "slightly nonlinear" one would systematically get the wrong causal direction under error misspecification. As optimizing Gaussian scores is the same as optimizing ℓ_2-loss, regressors that are more flexible than necessary for the causal model support anti-causal decisions. The intuition carries over when allowing for the flexibility of heteroskedastic error terms. If the true causal model has homoskedastic additive errors, fitting heteroskedastic models will increase the range of set-ups where misspecified Gaussian scoring chooses anti-causal relationships.

To overcome these issues, one could rely on general nonparametric independence tests, either between the different residuals or between residuals and predictors. Of course, this comes at higher computational cost and potentially lower sample efficiency in cases where Gaussian scoring works, including in the presence of non-monotonic causal effects.

5.A Proofs

5.A.1 Proof of Theorem 5.1

It is well known that for jointly Gaussian variables $X_1, \ldots, X_p$ every possible causal order can induce the multivariate distribution with a suitable linear model. As the sum of log-variances determines the Kullback-Leibler divergence in the Gaussian case, this sum must be equal for all these linear models that induce the same multivariate distribution.

For every possible multivariate distribution with existing and bounded moment matrix $\mathbf{\Sigma}^{\mathbf{X}} = \mathbb{E}\big[\mathbf{X}\mathbf{X}^\top\big]$, which the assumed model has, the linear least squares parameter and corresponding residual variances using arbitrary sets of regressor covariates are completely determined by the moment matrix. Therefore, the residual variances correspond to those of multivariate Gaussian data. Accordingly, one can obtain the same sum of log-variances as for the true model using the best linear predictors for every possible permutation. This proves the non-strict inequality in the theorem.

If for some variable j the conditional expectation given its parents (in the DAG G^π) is not a linear function, the linear model cannot be optimal in terms of residual variance. Therefore, $\sigma_j^{G^\pi}$ in an unrestricted model is lower than the residual standard error of the best fitting linear model, and the score is further improved. Hence, the inequality is strict as soon as there exists at least one such variable.

5.A.2 Proof of Proposition 5.1

The variances of X_1 and X_2 as well as $\mathcal{E}_2$ follow directly from the properties of the uniform distribution.

$$\left(\sigma_1^{G^0}\right)^2 = \mathrm{Var}(X_1) = \frac{1}{3}, \quad \left(\sigma_2^{G^0}\right)^2 = \mathrm{Var}(\mathcal{E}_2) = \frac{1}{3}, \quad \text{and}$$
$$\left(\sigma_2^{G^\perp}\right)^2 = \mathrm{Var}(X_2) = \beta^2 \mathrm{Var}(X_1) + \mathrm{Var}(\mathcal{E}_2) = \frac{1}{3}\big(\beta^2 + 1\big).$$

The last term requires some more work. Due to the symmetry, we can assume without loss of generality that $\beta \geq 0$. For the densities, we get

$$f_{X_1}(x_1) = \frac{1}{2}\mathbb{1}_{\{|x_1| \leq 1\}}, \quad f_{X_2|X_1}(x_2, x_1) = \frac{1}{2}\mathbb{1}_{\{|x_2 - \beta x_1| \leq 1\}} \quad \text{and}$$
$$f_{X_2}(x_2) = \int \frac{1}{4}\mathbb{1}_{\{|x_1| \leq 1\}}\mathbb{1}_{\{|x_2 - \beta x_1| \leq 1\}} dx_1 = \frac{1}{4}\left(\min\left\{1, \frac{x_2 + 1}{\beta}\right\} - \max\left\{-1, \frac{x_2 - 1}{\beta}\right\}\right) \coloneqq \frac{1}{4}(a - b).$$

For notational simplicity, we define random variables A and B with realizations a and b. We obtain the moments

$$\mathbb{E}[X_1|X_2] = \int x_1 f_{X_1|X_2}(x_1, X_2)dx_1 = \frac{\int \frac{x_1}{4}\mathbb{1}_{\{|x_1|\leq 1\}}\mathbb{1}_{\{|X_2-\beta x_1|\leq 1\}}dx_1}{\int \frac{1}{4}\mathbb{1}_{\{|x_1|\leq 1\}}\mathbb{1}_{\{|X_2-\beta x_1|\leq 1\}}dx_1} = \frac{A^2-B^2}{2(A-B)} = \frac{1}{2}(A+B),$$

$$\begin{aligned}\mathbb{E}\big[X_1^2|X_2\big] &= \int x_1^2 f_{X_1|X_2}(x_1, X_2)dx_1 = \frac{\int \frac{x_1^2}{4}\mathbb{1}_{\{|x_1|\leq 1\}}\mathbb{1}_{\{|X_2-\beta x_1|\leq 1\}}dx_1}{\int \frac{1}{4}\mathbb{1}_{\{|x_1|\leq 1\}}\mathbb{1}_{\{|X_2-\beta x_1|\leq 1\}}dx_1} = \frac{A^3-B^3}{3(A-B)} \\ &= \frac{1}{3}\big(A^2+AB+B^2\big),\end{aligned}$$

$$\mathrm{Var}(X_1|X_2) = \mathbb{E}\big[X_1^2|X_2\big] - \mathbb{E}[X_1|X_2]^2 = \frac{1}{12}(A-B)^2.$$

Finally, we are interested in

$$\begin{aligned}\left(\sigma_1^{G^\perp}\right)^2 &= \mathbb{E}[\mathrm{Var}(X_1|X_2)] = \mathbb{E}\left[\frac{1}{12}(A-B)^2\right] = \int_{-1-\beta}^{1+\beta}\frac{1}{12}(a-b)^2 f_{X_2}(x_2)dx_2 \\ &= \int_{-1-\beta}^{1+\beta}\frac{1}{48}(a-b)^3 dx_2 = 2\int_0^{1+\beta}\frac{1}{48}(a-b)^3 dx_2.\end{aligned}$$

Assume first $\beta \geq 1$. Then, $b = (x_2-1)/\beta$ in the area of integration. For $x_2 \geq \beta - 1$, it holds $a = 1$.

$$\begin{aligned}\left(\sigma_1^{G^\perp}\right)^2 &= \int_0^{1+\beta}\frac{1}{24}(a-b)^3 dx_2 = \int_0^{\beta-1}\frac{1}{24}\left(\frac{2}{\beta}\right)^3 dx_2 + \int_{\beta-1}^{1+\beta}\frac{1}{24}\left(\frac{\beta+1-x_2}{\beta}\right)^3 dx_2 \\ &= \frac{1}{24\beta^3}\left(8(\beta-1) + \int_0^2 u^3 du\right) = \frac{1}{24\beta^3}(8\beta-4) = \frac{1}{6\beta^3}(2\beta-1),\end{aligned}$$

where we applied the change of variable $u = \beta + 1 - x_2$ to simplify the integration. Inserting residual variances with $\gamma = \beta$, the proposition's statement follows.

Alternatively, if $\beta < 1$, $a = 1$ in the interval of integration. For $x_2 < 1 - \beta$, it holds $b = -1$.

$$\begin{aligned}\left(\sigma_1^{G^\perp}\right)^2 &= \int_0^{1+\beta}\frac{1}{24}(a-b)^3 dx_2 = \int_0^{1-\beta}\frac{1}{24}(2)^3 dx_2 + \int_{1-\beta}^{1+\beta}\frac{1}{24}\left(\frac{\beta+1-x_2}{\beta}\right)^3 dx_2 \\ &= \frac{1}{24}\left(8(1-\beta) + \frac{1}{\beta^3}\int_0^{2\beta} u^3 du\right) = \frac{1}{24}(8-4\beta) = \frac{1}{6}(2-\beta),\end{aligned}$$

where we applied the change of variable $u = \beta + 1 - x_2$ to simplify the integration. Inserting all the residual variances with $\gamma = 1/\beta$, the proposition's statement follows.

5.B Derivations for the figures

Assume model (5.4), which has model (5.3) as a special case for $\nu = 1$. With $X_1 \sim \mathcal{N}(0,1)$ the normalization is

$$V(\nu) \coloneqq \mathbb{E}\Big[|X_1|^{2\nu}\Big] = \frac{2^\nu \Gamma\left(\nu + \frac{1}{2}\right)}{\sqrt{\pi}}.$$

As before,

$$\left(\sigma_1^{G^0}\right)^2 = \mathrm{Var}(X_1), \quad \left(\sigma_2^{G^0}\right)^2 = \mathrm{Var}(\mathcal{E}_2) \quad \text{and}$$
$$\left(\sigma_2^{G^\perp}\right)^2 = \mathrm{Var}(X_2) = \beta^2 + \mathrm{Var}(\mathcal{E}_2).$$

5.B.1 Gaussian and uniform

For $\mathcal{E}_2 \sim \mathrm{Unif}[-1,1]$, X_1 is constrained to lie within

$$\mathrm{Sign}\Big((X_2-1)\sqrt{V(\nu)}/\beta\Big)\Big|(X_2-1)\sqrt{V(\nu)}/\beta\Big|^{1/\nu} \quad \text{and}$$
$$\mathrm{Sign}\Big((X_2+1)\sqrt{V(\nu)}/\beta\Big)\Big|(X_2+1)\sqrt{V(\nu)}/\beta\Big|^{1/\nu},$$

i.e., given X_2, it is a truncated standard Gaussian. The conditional mean and variance follow from standard theory about truncated Gaussian random variables. For simplicity, call the upper bound A and the lower bound B. The density of X_2 is then given by

$$f_2(x_2) = \int_b^a \frac{1}{2}\phi(x_1)dx_1 = \frac{1}{2}(\Phi(a) - \Phi(b)).$$

Finally,

$$\left(\sigma_1^{G^\perp}\right)^2 = \mathbb{E}[\mathrm{Var}(X_1|X_2)]$$

is obtained by numerically integrating over (a sufficient part of) the real line.

5.B.2 Two Gaussian random variables, or Gaussian and χ_1^2

Except for $V(\nu)$, all quantities are obtained by brute force numerical integration.

5.B.3 Two uniform random variables with heteroskedastic fitting

We can mainly follow the derivation in 5.A.2. Instead of $\left(\sigma_1^{G^\perp}\right)^2$, we need $\exp(\mathbb{E}[\log(\mathrm{Var}(X_1|X_2))])$ which is obtained by numerical integration.

Chapter 6

On the identifiability of causal location-scale noise models

Extracted from *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
International Conference on Machine Learning, PMLR 202, 14316-14332.

Abstract

We study the class of location-scale or heteroscedastic noise models (LSNMs), in which the effect Y can be written as a function of the cause X and a noise source N independent of X, which may be scaled by a positive function g over the cause, i.e., $Y = f(X) + g(X)N$. Despite the generality of the model class, we show the causal direction is identifiable up to some pathological cases.

6.1 Identifiability of LSNMs

In this section, we focus on the identifiability of location-scale noise models (LSNMs). A causal model is said to be identifiable under a set of structural constraints, if only the forward (causal) model is well specified and no backward model fulfilling these structural constraints exists.

To formally analyze this problem, we first need to define our assumed causal model.

Definition 6.1 (Location-Scale Noise Model)**.** Given two independent random variables X and N_Y. If the effect Y is generated by a location-scale noise model, we can express Y as an structural causal model (SCM) of the form

$$Y := f(X) + g(X)N_Y \ , \tag{6.1}$$

where $f\colon \mathcal{X} \to \mathbb{R}$ and $g\colon \mathcal{X} \to \mathbb{R}_+$, i.e. g is strictly positive.

LSNMs simplify to additive noise models (ANM) when $g(X)$ is constant, and to multiplicative noise models when $f(X)$ is constant.

To prove identifiability of such a restricted SCM, it is common to derive an ordinary differential equation (ODE), which needs to be fulfilled such that a backward model exists, see e.g. Hoyer et al. (2008a), or Zhang and Hyvärinen (2009). Intuitively, the solution space of such an ODE specifies all cases in which the model is non-identifiable, leaving all specifications which do not fulfill the ODE as identifiable. In the following theorem, we derive such a differential equation for LSNMs and discuss its implications.

Theorem 6.1. Assume the data is such that a location-scale noise model can be fit in both directions, i.e.,

$$\begin{aligned} Y &= f(X) + g(X)N_Y, \quad X \perp N_Y \\ X &= h(Y) + k(Y)N_X, \quad Y \perp N_X. \end{aligned}$$

Let $\nu_1(\cdot)$ and $\nu_2(\cdot)$ be the twice differentiable log densities of Y and N_X respectively. For compact notation, define

$$\begin{aligned} \nu_{X|Y}(x|y) &= \log\big(p_{X|Y}(x|y)\big) \\ &= \log\left(p_{N_X}\left(\frac{x - h(y)}{k(y)}\right)/k(y)\right) \\ &= \nu_2\left(\frac{x - h(y)}{k(y)}\right) - \log(k(y)) \quad \text{and} \\ G(x, y) &= g(x)f'(x) + g'(x)[y - f(x)]. \end{aligned}$$

Assume that $f(\cdot)$, $g(\cdot)$, $h(\cdot)$, and $k(\cdot)$ are twice differentiable. Then, the data generating mechanism must fulfill the following PDE for all x, y with $G(x, y) \neq 0$.

$$\begin{aligned} 0 = \nu_1''(y) + \frac{g'(x)}{G(x,y)}\nu_1'(y) + \frac{\partial^2}{\partial y^2}\nu_{X|Y}(x|y) + \\ \frac{g(x)}{G(x,y)}\frac{\partial^2}{\partial y \partial x}\nu_{X|Y}(x|y) + \frac{g'(x)}{G(x,y)}\frac{\partial}{\partial y}\nu_{X|Y}(x|y). \end{aligned} \tag{6.2}$$

The equality derived in Theorem 6.1 is equivalent to the result concurrently provided by Strobl and Lasko (2023) up to fixing a sign error in the terms involving $g'(x)$, which they had in a preliminary version. We derived this result independently using a different proof technique and additionally note that $p_{X|Y}(x|y)$ cannot be written as univariate function with argument $([x - h(y)]/k(y))$ if $Y \to X$ is an LSNM with non-constant $k(\cdot)$.

The conclusion of Strobl and Lasko (2023) is that if we have x_0 such that $G(x_0, y) \neq 0$ for all but countably many y, then knowing $\nu_{X|Y}(x_0|y)$, $G(x_0, y) \neq 0$, $g(x_0)$ and $g'(x_0)$ leads to $\nu_1(y)$ being constrained to a two dimensional affine space as (6.2) becomes an ODE. This is in analogy to the result of Hoyer et al. (2008a) for ANMs. For this case, Zhang and Hyvärinen (2009) have refined the result and provide a list of all possible cases of unidentifiable models: only for specific choices of $f(\cdot)$ and $\nu_2(\cdot)$, one can find $\nu_1(\cdot)$ such that the model is invertible.

This conclusion carries over to the LSNM. Assume there exist different values x such that $G(x, y) \neq 0$ for all but countably many y. If $g(\cdot)$ is strictly positive and $f(\cdot)$ is injective, this applies to all $x \in \mathbb{R}$ except for at most countably many. Each such value leads to a different ODE in y when plugging it into Equation (6.2). Only when the solution spaces of all ODEs overlap such that the same $\nu_1(\cdot)$ is found, which must also be valid log-density, the model can be invertible. This is not the case for generic combinations of $g(\cdot)$, $G(\cdot, \cdot)$ and $\nu_{X|Y}(\cdot|\cdot)$ but only for very specific exceptions. Thus, apart from some pathological cases, an LSNM cannot be invertible. A precise characterization of these cases as in Zhang and Hyvärinen (2009) for the post-nonlinear model, which involves the ANM as a special case, has not yet been found for LSNM to the best of our knowledge.

To provide a bit more intuition regarding the assumptions of Theorem 6.1, note that the results only apply to random variables X with unbounded support. This is implied by requiring that the log-density of N_X has to be twice differentiable. For example, X could not follow a uniform distribution. This also implies that $g(\cdot)$ has to be a non-linear (or constant) function since otherwise $g(\cdot)$ is negative for some attainable values of X and does not strictly map to $\mathbb{R}_+$ as required by our assumptions. Assuming that the noise variable is Gaussian, necessary conditions for the distributions of X and Y as well as the functions $f(\cdot)$, $g(\cdot)$, $h(\cdot)$, and $k(\cdot)$ can be found (Khemakhem et al., 2021). For completeness, we provide the corresponding result as Theorem 6.2 in Appendix 6.B.

6.A Proof

We follow the proof technique of Zhang and Hyvärinen (2009), i.e., we build upon the linear separability of the logarithm of the joint density of independent random variables. That is, for a set of independent random variables whose joint density is twice differentiable, the Hessian of the logarithm of their density function is diagonal everywhere (Lin, 1997). We first define the joint distribution $p(x, n_Y)$ via the change of variable formula, then derive the Hessian of its logarithm, and lastly, derive an PDE which is necessary to hold such that an inverse model can exist.

We define the change of variables from $\{x, n_Y\}$ to $\{y, n_X\}$

$$\begin{aligned} y &= f(x) + g(x)n_Y, \\ n_X &= [x - h(y)]/k(y). \end{aligned}$$

The according Jacobian matrix amounts to

$$\begin{pmatrix} \partial y/\partial x & g(x) \\ \dfrac{1}{k(y)} - \partial y/\partial x \dfrac{k(y)h'(y) + [x - h(y)]k'(y)}{k(y)^2} & -g(x)\dfrac{k(y)h'(y) + [x - h(y)]k'(y)}{k(y)^2} \end{pmatrix},$$

with absolute determinant $g(x)/k(y)$ such that

$$p(x, n_Y) = \frac{g(x)}{k(y)} p(y, n_X).$$

Under independence it holds

$$\begin{aligned} \frac{\partial^2}{\partial x \partial n_Y} \log(p(x, n_Y)) = 0 \quad \text{such that} \\ \frac{\partial^2}{\partial x \partial n_Y} \log\left(\frac{g(x)}{k(y)} p(y, n_X)\right) = \frac{\partial^2}{\partial x \partial n_Y} [\nu_1(y) + \nu_2(n_X) + \log(g(x)) - \log(k(y))] = 0. \end{aligned}$$

Evaluating this quantity and dividing by $(\partial y/\partial x)(\partial y/\partial n_Y)$ leads to

$$\begin{aligned} &\nu_1''(y) + \nu_1'(y)\left[\frac{g'(x)}{G(x, n_Y)}\right] + \nu_2''(n_X)\left[\frac{h'(y) + k'(y)n_X}{k(y)^2}\left[h'(y) + k'(y)n_X - \frac{g(x)}{G(x, n_Y)}\right]\right] + \\ &\nu_2'(n_X)\left[\frac{2\left[n_X k'(y)^2 + h'(y)k'(y)\right]}{k(y)^2} - \frac{h''(y) + n_X k''(y)}{k(y)} - \frac{h'(y)g'(x) + n_X k'(y)g'(x)}{k(y)G(x, n_Y)} - \right. \\ &\left. \frac{g(x)k'(y)}{k(y)^2 G(x, n_Y)}\right] + \frac{k'(y)^2 - k(y)k''(y)}{k(y)^2} - \frac{g'(x)k'(y)}{k(y)G(x, n_Y)} = 0, \end{aligned} \tag{6.3}$$

where

$$G(x, n_Y) = \frac{\partial y}{\partial x}\frac{\partial y}{\partial n_Y} = g(x)\big[f'(x) + g'(x)n_Y\big].$$

Assuming injectivity of $f(\cdot)$ and positivity of $g(\cdot)$, $G(x, n_Y)$ can only be 0 on a set of measure 0. Finally, plugging in all the definitions in Equation (6.2) and taking the derivatives, one finds that Equation (6.2) is identical to Equation (6.3).

6.B Gaussian noise

In the following, we state the theoretical result on Gaussian LSNMs found in the supplementary material in Khemakhem et al. (2021). Note that we slightly changed the theorem as the original version has a typo in the definition of g and k.

Theorem 6.2 (Khemakhem et al. (2021))**.** Assume the data follows the model in Def. 6.1 with N_Y standard Gaussian, $N_Y \sim \mathcal{N}(0, 1)$. If a backward model exists, i.e.

$$X = h(Y) + k(Y)N_X$$

where $N_X \sim \mathcal{N}(0, 1)$, $N_X \perp Y$ and $k > 0$, then one of the following scenarios must hold:

a) $(g, f) = \left(\frac{1}{\sqrt{Q}}, \frac{P}{Q}\right)$ and $(k, h) = \left(\frac{1}{\sqrt{Q'}}, \frac{P'}{Q'}\right)$ where Q, Q' are polynomials of degree two, $Q, Q' > 0$, P, P' are polynomials of degree two or less, and p_X, p_Y are strictly log-mix-rational-log. In particular, $\lim_{-\infty} g = \lim_{+\infty} g = 0^+$, $\lim_{-\infty} f = \lim_{+\infty} f < \infty$, similarly so for k, h, and f, g, h, k are not invertible.

b) g, k are constant, f, g are linear and p_X, p_Y are Gaussian densities.

Bibliography

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41(1):15–34.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

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

Azadkia, M., Chatterjee, S., and Matloff, N. (2021). *FOCI: Feature Ordering by Conditional Independence*. R package version 0.1.3.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the old faithful geyser. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(3):357–365.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Bowden, R. J. and Turkington, D. A. (1985). *Instrumental variables*. Cambridge University Press.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brockwell, P. J. and Davis, R. A. (2009). *Time series: theory and methods*. Springer science & business media.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

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

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., Zhao, L., et al. (2019a). Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., Zhao, L., et al. (2019b). Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565.

Ćevid, D., Bühlmann, P., and Meinshausen, N. (2020). Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2021). *xgboost: Extreme Gradient Boosting*. R package version 1.4.1.1.

Chevalley, M., Roohani, Y., Mehrjou, A., Leskovec, J., and Schwab, P. (2023). Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*.

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

Davidson, J. (2002). Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes. *Journal of Econometrics*, 106(2):243–269.

Davidson, J. and de Jong, R. (1997). Strong laws of large numbers for dependent heterogeneous processes: a synthesis of recent and new results. *Econometric Reviews*, 16(3):251–279.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4):533–558.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5):507–534.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021). Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778.

Godfrey, L. G. (1989). *Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches*. Cambridge University Press.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guo, Z., Ćevid, D., and Bühlmann, P. (2022). Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, 50(3):1320.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 46(6):1251–1271.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008a). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., and Shimizu, S. (2008b). Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 282–289.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5).

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Immer, A., Schultheiss, C., Vogt, J. E., Schölkopf, B., Bühlmann, P., and Marx, A. (2023). On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Khemakhem, I., Monti, R., Leech, R., and Hyvarinen, A. (2021). Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 3520–3528. PMLR.

Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S., and Qi, L. S. (2013). Crispr interference (crispri) for sequence-specific control of gene expression. *Nature Protocols*, 8(11):2180–2196.

Lehmann, E. L., Romano, J. P., and Casella, G. (2005). *Testing statistical hypotheses*. Springer, 3 edition.

Lin, J. (1997). Factorizing multivariate function classes. In *Advances in Neural Information Processing Systems*, volume 10, pages 563 – 569.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Maddala, G. and Lahiri, K. (2009). *Introduction to Econometrics*. John Wiley & Sons Ltd, 4th edition.

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

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

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

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368.

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

Mooij, J. M. and Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 431–439.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(1):1103–1204.

Nowzohour, C. and Bühlmann, P. (2016). Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485.

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

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

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

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

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

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

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, volume 26.

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

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053.

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

Pfister, N. and Peters, J. (2019). *dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion*. R package version 2.1.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.

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

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Schultheiss, C. and Bühlmann, P. (2023a). Ancestor regression in linear structural equation models. *Biometrika*, 110(4):1117–1124.

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

Schultheiss, C. and Bühlmann, P. (2024a). Ancestor regression in structural vector autoregressive models.

Schultheiss, C. and Bühlmann, P. (2024b). Assessing the overall and partial causal well-specification of nonlinear additive noise models. *Journal of Machine Learning Research*, 25(159):1–41.

Schultheiss, C., Bühlmann, P., and Yuan, M. (2024). Higher-order least squares: assessing partial goodness of fit of linear causal models. *Journal of the American Statistical Association*, 119(546):1019–1031.

Shah, R. D. and Bühlmann, P. (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):113–135.

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

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

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030.

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

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*, chapter 5.4.2. MIT Press.

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

Taeb, A., Gamella, J. L., Heinze-Deml, C., and Bühlmann, P. (2023). Learning and scoring gaussian latent variable causal models with unknown additive interventions. *arXiv preprint arXiv:2101.06950*.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, 4th edition.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. (2023). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072.

Wang, Y. S. and Drton, M. (2020). High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59.

Wood, S. N. (2011). 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.

Xu, S., Mian, O. A., Marx, A., and Vreeken, J. (2022). Inferring cause and effect in the presence of heteroscedastic noise. In *International Conference on Machine Learning*, pages 24615–24630. PMLR.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, J. and Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655.