Diss. ETH No. 23806

CAUSAL INFERENCE IN SEMIPARAMETRIC AND NONPARAMETRIC STRUCTURAL EQUATION MODELS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

JAN ERNEST

MSc ETH Mathematics
born on 09.04.1987
citizen of Zürich ZH

accepted on the recommendation of

Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Marloes M. Maathuis, co-examiner

2016

Felix qui potuit rerum cognoscere causas

– Virgil, Georgica, book 2, verse 490, 29 BCE

Acknowledgments

First and foremost, I want to express my deepest gratitude to my advisor Peter Bühlmann for his continuous encouragement and enthusiasm throughout my PhD. He gave me excellent advice and provided exceptional support over all the years. His statistical intuition is second to none and every discussion with him provided me with new confidence, inspiration and dozens of new approaches. Peter gave me a lot of freedom in the choice of my projects and motivated me to pursue my own ideas. Moreover, he always made sure to maintain a fantastic work environment with an excellent work-life-balance.

Many thanks go to Marloes Maathuis for fruitful discussions about causal inference and for acting as my co-examiner for the thesis.

Second, but not less important, I want to thank all the current and former members of the Seminar for Statistics for creating such a stimulating, familial atmosphere and for making my PhD an unforgettable time. Special thanks to Susan for being so helpful; to Jonas, Dominik and Shu for inspiring and amicable collaborations; to my office mates Michaël, Alan and Marco for interesting on- and off-topic discussions and for making my time in the office so enjoyable; and to Anna, Patric, Ema, Gian, Preetam, Ruben, Nicolas and all other friends and colleagues for all the valuable discussions about research and the countless fun activities.

Lastly, I wish to express my sincere gratitude to my wonderful wife Stephanie and to my parents for their unconditional and never-ending support.

Contents

Abstract

The goals of causal inference are inherently different from the ones of classical statistics. Instead of measuring statistical associations between variables, the main focus is on the characterization of the underlying causal mechanisms. This is typically achieved via the estimation of causal graphs (*structure learning*) or the prediction of causal effects under interventions. Both problems are well-understood and elaborated for linear structural equation models (SEMs). This thesis addresses them for specific classes of semiparametric and nonparametric SEMs.

First, we study structure learning for causal additive models (CAMs). CAMs constitute a natural semiparametric extension of linear Gaussian SEMs: while still relying on the additivity of the functions and Gaussianity of the noise, all functions are assumed to be exclusively nonlinear. We present a score-based structure learning methodology based on (restricted) maximum likelihood estimation that is consistent in low- and high-dimensional settings. The key idea of our approach is to decouple order search among the variables from subsequent edge selection in the graph. We provide an efficient implementation of our proposed methodology in the `R`-package `CAM` and evaluate its performance in extensive simulations.

In the second part of the thesis, we study the identifiability and estimation of partially linear additive SEMs with Gaussian noise (PLSEMs). Thus, we drop the assumption of exclusivity of the functional type and with that we address one of the major limitations of both, linear SEMs and CAMs. We precisely specify how linear and nonlinear additive functions impose restrictions on the underlying causal model and derive a systematic characterization of the identifiability of PLSEMs. Thereby, we close a relevant gap, as the identifiability theory of additive models with Gaussian noise was only elaborated for linear SEMs and CAMs. We complement

the theoretical findings with an efficient score-based estimation procedure that, given one PLSEM, finds all equivalent PLSEMs. We prove low- and high-dimensional consistency results for our algorithm and evaluate its performance on simulated datasets.

In the last part, we additionally relax the additivity and Gaussianity assumptions. Structure learning for unstructured nonparametric SEMs is a highly ambitious task as it is plagued by the curse of dimensionality. Interestingly, the situation is different for the estimation of (total) causal effects. We show that a specific marginal integration regression technique (S-mint) theoretically achieves the optimal univariate convergence rate of nonparametric regression for a very general class of nonparametric SEMs with known (or approximately known) structure (assuming sufficient smoothness). Specifically, S-mint does not suffer from the curse of dimensionality. We propose an implementation based on an additive regression approximation with subsequent L_2-boosting. In extensive simulations, our method demonstrates a more pronounced robustness with respect to model misspecification than other methods that rely more heavily on the correct estimation of the causal structure.

Zusammenfassung

Das Gebiet der kausalen Inferenz hat eine grundsätzlich andere Zielsetzung als die klassische Statistik. Statt statistische Assoziationen zwischen Variablen zu messen, besteht der Hauptfokus der kausalen Inferenz darin, die zugrundeliegenden kausalen Zusammenhänge zu charakterisieren. Dies kann auf verschiedene Arten angegangen werden. Zum Beispiel, indem man einen Graphen schätzt, der die kausalen Mechanismen abbildet. Alternativ kann man versuchen, direkt die kausalen Effekte vorherzusagen, welche durch Interventionen verursacht werden. Beide Ansätze sind seit längerem bekannt und ausgereift für lineare Strukturgleichungsmodelle. Die vorliegende Doktorarbeit untersucht diese Ansätze in spezifischen Klassen von semiparametrischen und nichtparametrischen Strukturgleichungsmodellen.

Eine naheliegende semiparametrische Erweiterung der linearen Gauss'schen Modelle sind sogenannte kausale additive Modelle. Sie sind immer noch additiv mit Gauss'schen Fehlertermen, bestehen jedoch ausschliesslich aus nichtlinearen Funktionen. Wir entwickeln eine score-basierte Maximum-Likelihood Methode, um für diese Modellklasse die zugrundeliegenden kausalen Graphen zu schätzen und zeigen deren Konsistenz für den tief- und hoch-dimensionalen Fall. Die entscheidende Idee der Methode besteht darin, die Suche nach einer korrekten kausalen Ordnung der Variablen von der Suche nach individuellen Kanten im kausalen Graphen zu entkoppeln. Wir stellen eine effiziente Implementierung der Methode im `R`-Paket `CAM` zur Verfügung und untersuchen deren Leistungsfähigkeit in diversen numerischen Experimenten.

Sowohl die linearen Gauss'schen Modelle als auch die kausalen additiven Modelle besitzen den grossen Nachteil, dass alle additiven Komponenten vom selben Typ sein müssen, das heisst, entweder alle linear oder alle nichtlinear. Diese restriktive Annahme kann umgangen werden, indem man

beide Funktionstypen im gleichen Modell zulässt, das heisst, durch Betrachten von partiell linearen additiven Strukturgleichungsmodellen mit Gauss'schen Fehlertermen. Wir untersuchen, wie lineare und nichtlineare Funktionen das zugrundeliegende kausale Modell einschränken, und leiten daraus eine systematische Charakterisierung der Identifizierbarkeit der gesamten Modellklasse her. Dadurch schliessen wir eine grosse Lücke in der Identifizierbarkeitstheorie additiver Modelle, welche bisher nur für lineare Gauss'sche und kausale additive Modelle ausgearbeitet wurde. Wir ergänzen die Theorie durch einen effizienten Algorithmus, der für ein gegebenes partiell lineares Modell alle dazu äquivalenten Modelle auflistet. Wir beweisen dessen Konsistenz im tief- und hoch-dimensionalen Fall und untersuchen die Leistungsfähigkeit auf simulierten Datensätzen.

Zuguterletzt stellen wir uns die Frage, welche Aussagen ohne die Annahme von additiven Funktionen und Gauss'schen Fehlertermen getroffen werden können. Unglücklicherweise ist das Schätzen von unstrukturierten nichtparametrischen Modellen geprägt vom sogenannten Fluch der Dimensionalität. Dies ist interessanterweise nicht der Fall, wenn wir versuchen, (totale) kausale Effekte zu schätzen. Wir zeigen, dass eine spezifische Regressionsmethode, die auf marginaler Integration beruht, für eine allgemeine Klasse von nichtparametrischen Strukturgleichungsmodellen mit bekannter (oder ungefähr bekannter) Struktur die optimale univariate Konvergenzrate für nichtparametrische Regression erreicht (unter Annahme genügender Differenzierbarkeit). Insbesondere umgeht diese Methode den Fluch der Dimensionalität. Als Ergänzung zur Theorie schlagen wir eine Implementierung der Methode vor, welche auf einer additiven Approximation mit anschliessendem L_2-boosting beruht. In ausgiebigen Simulationen erweist sich diese Methode als robuster gegenüber Abweichungen vom Modell als andere Schätzverfahren, welche stärker von der korrekten Schätzung der kausalen Struktur abhängig sind.

Nomenclature

Abbreviations

RCT	randomized controlled trial
DAG	directed acyclic graph
PDAG	partially directed acyclic graph
CPDAG	completed partially directed acyclic graph
SEM	structural equation model
PLSEM	partially linear structural equation model
CAM	causal additive model
MLE	maximum likelihood estimation/estimator
SHD	structural Hamming distance
SID	structural intervention distance
GES	greedy equivalence search
PC	PC-algorithm
CPC	conservative PC-algorithm
LiNGAM	Linear non-Gaussian acyclic model
RESIT	regression with subsequent independence tests
S-mint	marginal integration with adjustment set S
est S-mint	*S-mint* based on estimated graph or order
IPW	inverse probability weighting
DR	double robust
TMLE	targeted maximum likelihood estimation
BIC	Bayesian information criterion
CPU	central processing unit
fMRI	functional magnetic resonance imaging

Symbols

p / n	number of variables / number of samples
$X = (X_1, \ldots, X_p)$	observed system of random variables
$X^{(1)}, \ldots, X^{(n)}$	n i.i.d. copies of $X \in \mathbb{R}^p$
$D = (V, E)$ / G	DAG / PDAG
$V = \{1, ..., p\}$	vertex set of DAG
$E \subset V^2$	edge set of DAG
$i, j, k, l \in V$	nodes in a DAG (corresp. to X_i, X_j, X_k, X_l)
j_X, j_Y	indices of variables X and Y (e.g., $X_{j_Y} = Y$)
$\deg_D(i)$	degree of node i in DAG D
$i \to j,\ (i, j)$	directed edge in a graph
$i - j$	undirected edge in a graph
μ_j	intercepts in a SEM
f_j / $f_{j,i}$	non-additive / additive functions in a SEM
ε_j	noise variables in SEM
σ_j^2	variances of noise variables in SEM
$\mathcal{F}, \mathcal{F}_i, \mathcal{F}^{\oplus\ell}/\mathcal{F}_n, \mathcal{F}_n^{\oplus\ell}$	function spaces / approximation spaces
$P, \mathbb{P}$	distribution of $X_1, \ldots, X_p$
$P_n, \mathbb{P}_n$	empirical distribution of $X^{(1)}, \ldots, X^{(p)}$
π, σ	permutations on $\{1, ..., p\}$
X^π	permuted random variables: $X_j^\pi = X_{\pi(j)}$
ξ_p	min. degree of separation of true and wrong models
$\mathrm{pa}_D(j), \mathrm{pa}(j)$	parents of node j in DAG D
$\mathrm{do}(X = x)$	do-intervention on variable X
$S, S(j_X)$	valid backdoor adjustment set for variable X
K, L, h_1, h_2	kernel functions and corresponding bandwidths
$CS_{k\to j}^{\mathrm{rel}}$	relative causal strength of edge $k \to j$
$F,\ \mathrm{D}F$	PLSEM-function / its Jacobian
$\mathscr{D}(\mathbb{P})$	distribution equivalence class
$G_{\mathscr{D}(\mathbb{P})}$	graphical PDAG representation of $\mathscr{D}(\mathbb{P})$
$\mathcal{F}(\mathbb{P})$	set of potential PLSEM-functions
$\mathcal{S}(\mathbb{P})$	set of potential causal orderings
$\mathcal{V}$	set of restricted index tuples
$\mathcal{C}^2(\mathbb{R})$	set of twice continuously differentiable functions
L_2	set of square-integrable functions
P	pattern of DAG D
$\mathcal{K}$	consistent set of background knowledge
$G_{P,\mathcal{K}}$	maximally oriented PDAG with respect to P and $\mathcal{K}$

Chapter 1

Introduction

In this chapter we first give a general introduction to the field of causal inference from observational data and discuss its main goals. Next, we explain the key concepts behind directed acyclic graph models and structural equation models, and question the commonly used assumptions. Finally, we outline the scope of this thesis, assess the novelty of our contributions and explain how they relate to the existing framework.

1.1 Causal inference from observational data

> *The research questions that motivate most quantitative studies in the health, social and behavioral sciences are not statistical but causal in nature.*
>
> – Judea Pearl, The Science and Ethics of Causal Modeling, 2010

The wish to establish cause-effect relations is omnipresent in science and everyday life. A researcher may want to investigate new genetic causes of cancer; a pharmaceutical company may want to assess the efficacy of a new sleeping pill; a car insurance company may be interested in predicting the reduction of the number of accidents if they enforce the installation of drive recorders in all the policyholders' cars; or a person lying awake in bed may wonder whether he would be sleeping now if he had not drunk two cups of coffee after dinner.

Establishing cause-effect relations is fundamentally different from finding statistical associations. Nevertheless, these two concepts are frequently mixed up (which is referred to as the *correlation-implies-causation-fallacy*). Suppose that a news article proclaims: "Employees with grey hair earn more money". Does that mean that dyeing my hair grey will make my salary increase? The answer is: maybe. While the statement in the news reports an association (a co-occurrence of grey hair and a higher salary), we tried to draw the conclusion that grey hair is a cause of a higher salary. This may, indeed, be true: grey hair may make us look more experienced and knowledgeable and that could cause an increase in salary. However, it could also be the opposite, namely, that people who earn more money are under higher pressure to perform which causes their hair to turn grey. Or, it may be the case that grey hair and a high salary are not causally related at all. There may be a third confounding variable that affects both, grey hair and a high salary, that is responsible for the observed association. For example, the age of the employees. So how can we specifically address a causal question?

Randomized controlled trials

The gold-standard is to conduct a randomized controlled trial (RCT). Consider the example of the pharmaceutical company that wants to assess the efficacy of a new sleeping pill. In an RCT, participants are randomly selected and randomly assigned to either the treatment group (receiving the new sleeping pill) or the control group (receiving a placebo). The randomization (ideally) controls for the effects of confounding variables (such as the age or previous medical history of the participants), and the only difference between the groups lies in the type of treatment. This allows us to draw conclusions about the (unconfounded) effect of the treatment (sleeping pill versus placebo) directly.

Undoubtedly, having experimental data from an RCT that is designed specifically for answering the causal question of interest is one of the best scenarios for causal inference. So why is it not always possible to rely on it?

The need for approaches different from randomization

Practically, there are many scenarios in which it is impossible to perform suitable RCTs to test the causal hypotheses at hand. The most common

example is the one where performing an RCT would be unethical. Suppose we want to address the question whether obesity causes heart disease. In an RCT, we would randomly split the participants of the given population into two groups and constrain the participants of one group to become obese. For obvious reasons, it is rather unethical to expose a randomly selected subset of our population to a risk factor in order to determine whether or not it is a cause of the disease.

Another major problem that can arise in many scientific disciplines is that the sheer number of causal hypotheses is simply too large to test experimentally. Good examples are gene knockout experiments, where one wants to assess the function of particular genes (on an outcome of interest) by intervening on these genes and rendering them inoperative. In these intervention experiments, the number of potential candidate genes for single knockout experiments typically lies in the thousands, not to mention the number of potential experiments when also allowing for simultaneous knockouts of several genes. In this situation, it can be too time-consuming or too expensive to perform all these experiments. The ability to predict *a priori* which (combinations) of the intervention targets are most likely to have a strong effect on the outcome of interest would be tremendously useful for the design and prioritization of knockout experiments.

Observational data

In this thesis, we focus on the setting where experimental data from RCTs is not available. That means, we have to rely on observational data to answer causal questions. As the name implies, observational data is obtained by pure observation of a system of interest without subjecting it to any kind of external manipulations. The good news is that in most applications, observational data is either readily available or quite cheap to collect.

Inferring causal relations from observational data is one of the most traditional areas of causal inference and many different concepts and frameworks were established over the last decades. Examples include structural equation modeling (cf. Bollen, 1998), the theory of potential outcomes and counterfactuals (cf. Dawid, 2000; Rubin, 2005) or the use of instrumental variables (cf. Angrist et al., 1996; Didelez et al., 2010).

More recent approaches try to exploit heterogeneity in the data and to account for all kinds of additional sources of data (e.g., interventional

data) or background information (e.g., expert or prior knowledge). Examples are the greedy interventional equivalence search (GIES) (Hauser and Bühlmann, 2012), the method of invariant prediction (Peters et al., 2015) or constraint-based optimization techniques encoding causal constraints as SAT instances (cf. Hyttinen et al., 2013, 2014; Triantafillou and Tsamardinos, 2015; Triantafilou et al., 2010).

Causal sufficiency

Throughout the thesis, we make the assumption that our observed system of variables satisfies causal sufficiency. This means that we do not allow for any unknown (hidden) common causes of any of the observed variables.

Let us go back to the example where we want to assess whether obesity causes heart disease. To establish a causal relation, we have to rule out that the observed association between obesity and heart disease is spurious, that is, only due to a common cause. The assumption of causal sufficiency now requires that all possibly relevant common causes are measured. It is highly questionable whether we can think of all potential common causes of obesity and heart disease and measure all of them. Hence, the assumption of causal sufficiency is highly unlikely to hold in this example. Remedial action is taken by approaches that address causal inference in the presence of hidden variables, see Colombo et al. (2012), Maathuis and Colombo (2015), Perković et al. (2016), Richardson and Spirtes (2002), Shpitser et al. (2011), Spirtes et al. (2000), and Zhang (2008) and references therein.

Realistically, the assumption of causal sufficiency is probably violated in most practical scenarios. Still, there is a legitimate hope that in some cases, the influence of unmeasured variables is relatively small.

1.2 Directed acyclic graph models

In the area of causal inference from observational data one commonly assumes that the observed system of variables is driven by an underlying causal mechanism and typically pursues the aim of either recovering (parts of) that causal mechanism (*structure learning*) or predicting the response to external manipulations of the system (*estimation of causal effects under interventions*).

Throughout the thesis, we assume that these underlying causal mechanisms can be represented in terms of directed acyclic graphs (DAGs). We give an illustrative example of a DAG in Figure 1.1, precise definitions can be found in Section 3.1.1.

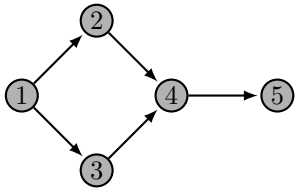


Figure 1.1: Example of a DAG. All edges are directed and there are no directed cycles. The nodes $1, \ldots, 5$ correspond to the observed random variables $X_1, ..., X_5$.

The use of directed graphs over undirected graphs has the natural advantage that edge orientations come along with an intuitive causal interpretation in the sense that an edge $2 \to 4$ in the DAG reflects a (direct) causal influence of X_2 (*cause*) on X_4 (*effect*). This means that if we intervene on the variable X_2, the effect of this manipulation will be propagated according to the direction of the edges and affect the variable X_4 (and also X_5, but not X_1 and X_3). In case of an intervention on X_4, the variable X_2 remains unaffected. In Section 1.3, we introduce the concept of structural equation models to formalize this notion.

The assumption of acyclicity of the causal structure is quite standard, but comes with the limitation that our model doesn't allow for the incorporation of any sort of feedback mechanisms, which are widespread, especially in biological systems. In principle, the assumption can be justified by arguing that the underlying driving mechanisms are indeed acyclic, when the system of variables is observed in a sufficiently fine time interval.

Models that account for feedback mechanisms and address cyclic structures are discussed in Lacerda et al. (2008), Mooij et al. (2011), Mooij and Heskes (2013), Richardson (1996), Rothenhäusler et al. (2015), and Spirtes (1995) and references therein.

Causal Markov condition

DAGs are most commonly used to represent conditional independences in a given distribution. This can be done via the causal Markov condition,

which requires that every variable is independent of its non-descendants in the DAG (excluding the parents) given its parents in the DAG (cf. Spirtes et al., 2000, Sections 3.4.1 and 3.5.1). As an example, the causal Markov condition applied to the DAG in Figure 1.1 encodes the (non-trivial) conditional independences $X_2 \perp\!\!\!\perp X_3 \mid X_1$, $X_4 \perp\!\!\!\perp X_1 \mid \{X_2, X_3\}$ and $X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\} \mid X_4$.

The causal Markov condition allows us to relate distributional properties of the variables to graphical properties of the DAG. For example, the joint distribution factorizes according to the DAG structure (Lauritzen, 1996, Theorem 3.27). Also, the DAG encodes conditional independences in the distribution via the criterion of d-separation. In words, the d-separation criterion characterizes how to "read off" conditional independences from specific edge constellations in the DAG, see Section 1.2.3 in Pearl (2009) for precise definitions.

Suppose that we have observational data and we want to address structure learning for DAG models, or even simpler, find out whether two specific variables in the system are causally related. In general, the causal Markov condition alone is not enough to help us address that problem. By definition, many different graphs satisfy the causal Markov condition with respect to the given distribution. In particular, all complete DAGs (DAGs in which any pair of nodes is connected by an edge), as they do not encode any non-trivial conditional independences in the distribution via the causal Markov condition. This implies that for any pair of variables X and Y it is not possible to draw conclusions about their causal relation only based on the causal Markov condition. There always exist two DAGs that both satisfy the causal Markov condition with respect to the observed distribution, such that X causes Y in one DAG and Y causes X in the other DAG.

Faithfulness condition

One common approach to address causal structure learning from observational data is to assume the faithfulness condition in addition to the causal Markov condition. It requires that every conditional independence that holds in the distribution must be entailed by the causal Markov condition applied to the DAG (cf. Spirtes et al., 2000, Sections 3.4.3 and 3.5.2). In particular, the faithfulness condition enforces a one-to-one correspondence between conditional independences in the distribution and d-separation statements in the DAG.

While the causal Markov condition is widely accepted, the (generally untestable) faithfulness condition is more often criticized. For a detailed discussion, see Zhang and Spirtes (2008) and references therein. Throughout the thesis, we try not to assume the faithfulness condition whenever possible. We only rely on it in Sections 3.2.1 and 3.3 to prove similarities of our theoretical approach to earlier results in the field.

Markov equivalence classes

The space of DAGs clusters in Markov equivalence classes, which consist of all DAGs that satisfy the same set of d-separation statements (and hence entail the same conditional independence relations in the distribution via the causal Markov condition). Markov equivalence classes are well-understood and various characterizations of them exist. For example, all DAGs in a Markov equivalence class share the same skeleton and v-structures (Verma and Pearl, 1990), can be compactly represented by a single partially directed acyclic graph (cf. Andersson et al., 1997; Chickering, 2002; Meek, 1995), or can be transformed into each other via sequences of covered edge reversals (Chickering, 1995). We discuss these results in detail in Chapter 3.

Under the additional assumption of faithfulness, the conditional independences in the distribution of the variables (due to the one-to-one correspondence with d-separation statements in the graph) precisely characterize the Markov equivalence class of the true underlying DAG (as all these DAGs satisfy the same set of d-separation statements). In this case, we say that the Markov equivalence class of the underlying DAG is *identifiable* from the distribution. Many common structure learning methods rely on the faithfulness condition. A well-known example is the PC-algorithm (Spirtes and Glymour, 1991), which seeks to infer the Markov equivalence class of the underlying DAG from conditional independences in the distribution.

Without making additional assumptions on the data-generating process, the DAGs in a Markov equivalence class cannot be further distinguished based on the properties of the distribution. This is where structural equation models play a crucial role, as they are a means of putting specific restrictions on the (causal) data-generating process.

1.3 Structural equation models

> *... Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? These are causal questions because they require some knowledge of the data-generating process;*
>
> – Judea Pearl, The Science and Ethics of Causal Modeling, 2010

Throughout the thesis, we encode our "knowledge of the data-generating process" via the assumption that the joint distribution of the variables has been generated by a structural equation model (SEM) with an underlying DAG structure (which we will also refer to as the *associated* or *corresponding* DAG). A SEM functionally relates the marginal distribution of each variable to the distribution of its direct effects (the parents in the associated DAG) and random noise. For a detailed description of the general model and precise assumptions, we refer the reader to Section 3.1.1.

SEMs naturally relate to the previously discussed framework. In fact, SEMs have the nice property that, by construction, the distribution generated by them satisfies the Markov factorization property and the causal Markov condition with respect to the associated DAG, see Theorem 1.4.1 in Pearl (2009) and the related discussion.

The SEM corresponding to our model-DAG in Figure 1.1 is given as

$$\begin{aligned} X_1 &= f_1(\varepsilon_1) \\ X_2 &= f_2(X_1, \varepsilon_2) \\ X_3 &= f_3(X_1, \varepsilon_3) \\ X_4 &= f_4(X_2, X_3, \varepsilon_4) \\ X_5 &= f_5(X_4, \varepsilon_5), \end{aligned} \tag{1.1}$$

where $\varepsilon_1, \ldots, \varepsilon_5$ are the noise variables, which in our context are always mutually independent due to the assumption of causal sufficiency. The functions $f_1, \ldots, f_5$ represent the causal mechanisms underlying the system, which makes SEMs inherently asymmetric and their interpretation different from classical mathematical equations. Notably, the variables on the right hand side of the equal sign are considered to be direct causes of the variables on the left hand side. We sometimes write arrows ($\leftarrow$)

instead of equal signs to account for this interpretation. The associated DAG can easily be recovered from the SEM by drawing directed edges from the variables on the right hand side of the equation to the variables on the left hand side.

The generality of SEMs

Regarding the task of causal structure learning, SEMs, in their most general form, face the same limitation as the DAG models. Suppose we only know the distribution generated by SEM (1.1) and want to draw conclusions about the underlying causal structure (the associated DAG in Figure 1.1). Without additional restrictions or assumptions, SEMs are simply too general to do that. Recall from Section 1.2 that every distribution satisfies the causal Markov condition with respect to all complete DAGs. In fact, one can easily construct corresponding SEMs that generate the given distribution for all these complete DAGs (Peters et al., 2014, Proposition 9).

Restricted SEMs

A recent approach to achieve better identifiability and estimation properties is to consider specific classes of restricted SEMs, where one puts restrictions on the functions (f_j), the noise variables (ε_j), or both.

The main research problems that are addressed for restricted SEMs can typically be categorized into the following three distinct (but closely related) tasks:

1. *Identifiability of the restricted SEM and the associated DAG.*
   Given an infinite sample from a distribution generated by a specific class of restricted SEM: can we recover the true data-generating SEM and its associated DAG from the distribution? If not, do all the SEMs (and their associated DAGs) that generate the same distribution share certain structural properties? For example, given a specific pair of variables, is there a causal relation between them and is it the same in all the SEMs?

2. *Structure learning of the associated DAG.*
   Given a finite sample from a distribution generated by a specific class of restricted SEM: can we design an estimation methodology to

learn the associated DAG (or the set of all possible DAGs) from the distribution? What kind of properties does the methodology have? For example, is it consistent in low- or high-dimensional settings?

3. *Estimation of causal effects under interventions.*
   Given a finite or infinite sample from a distribution generated by a specific type of restricted SEM and additionally assuming that the underlying DAG (or the set of all possible underlying DAGs) is known: how can we estimate the causal effect of an intervention on one or several of the variables on a response variable of interest?

Until recently, these three problems were mainly studied for linear SEMs, as the assumption of linearity of the functions entails good estimation properties. The identifiability and structure learning properties of linear SEMs crucially depend on the assumption on the noise distributions: see Chickering (2002), Kalisch and Bühlmann (2007), Nandy et al. (2016), Spirtes et al. (2000), and Spirtes and Zhang (2016) for the (most common) case of Gaussian noise; Shimizu et al. (2006) and Shimizu et al. (2011) for non-Gaussian noise; and Hoyer et al. (2008) for arbitrary noise distributions. An in-depth discussion of these references is given in Section 3.1.2. Pearl (2009) gives a broad overview of the framework of do-calculus for the estimation of causal effects under interventions; see also Section 4.1.1 for a brief introduction.

Unfortunately, the exclusive linearity assumption is restrictive and at best approximately true in most practical situations. This brings up the question of what can be done in more general classes of restricted SEMs.

1.4 Scope of this thesis

The central theme of this cumulative dissertation is the study of specific classes of semiparametric and nonparametric extensions of linear SEMs. More precisely, we examine combinations of the following types of restrictions:

ADDITIVITY	The functions f_j are additive in all arguments.
EXCLUSIVITY	The functions f_j are of exclusively nonlinear form.
GAUSSIANITY	Gaussianity of noise: $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j^2 > 0$.

On our way through the chapters, we relax more and more of the assumptions involved and increase the generality of the considered classes of restricted SEMs.

Chapter 2: Structure learning for causal additive models

EXCLUSIVITY ADDITIVITY GAUSSIANITY

In Chapter 2, we consider structure learning for a recently proposed model class that relies on all three restrictions, which we denote as *causal additive models (CAMs)*. For the DAG in Figure 1.1, the corresponding CAM is of the form

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= f_{2,1}(X_1) + \varepsilon_2 \\ X_3 &= f_{3,1}(X_1) + \varepsilon_3 \\ X_4 &= f_{4,2}(X_2) + f_{4,3}(X_3) + \varepsilon_4 \\ X_5 &= f_{5,4}(X_4) + \varepsilon_5, \end{aligned} \tag{1.2}$$

where all $f_{j,i}(\cdot) : \mathbb{R} \to \mathbb{R}$ are assumed to be nonlinear and three times differentiable functions and $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $\sigma_j^2 > 0$. With the additivity of the functions and the Gaussianity of the noise variables, CAMs constitute a natural structured semiparametric extension of linear Gaussian SEMs. In terms of identifiability, however, they are crucially different: while the associated DAG of a linear Gaussian SEM is only identifiable up to a Markov equivalence class under the assumption of the faithfulness condition (see, e.g., Spirtes and Zhang, 2016), it is fully identifiable in CAMs even without assuming the faithfulness condition. This result was first presented as a corollary of the general identifiability result of additive

noise models in Peters et al. (2014) and is restated in Lemma 1 (due to its importance).

In Chapter 2, we address structure learning of the (unique) associated DAG for CAMs. We develop a score-based (restricted) maximum likelihood methodology that is consistent in the low- and high-dimensional settings (assuming sufficient sparsity). The treatment of the high-dimensional scenario with a restricted maximum likelihood approach is novel.

We additionally propose an algorithm to efficiently estimate the associated DAG based on a greedy search strategy. We evaluate its performance in various experiments on simulated and real data. It is the first algorithm that addresses structure learning for low- and high-dimensional CAMs.

Chapter 3: Identifiability & estimation of partially linear SEMs

As a motivation for the search of generalizations of linear SEMs we questioned the validity of the linearity assumption in practical situations. Even though CAMs are more general, the same criticism can be brought up for them, as they rely on the assumption of exclusively nonlinear functions.

In Chapter 3, we relax the assumption of exclusivity of the functional type and consider partially linear additive structural equation models with Gaussian noise (PLSEMs). As an example, the PLSEM corresponding to the DAG in Figure 1.1 is given by equation (1.2) except that we drop the nonlinearity assumption on the functions, that is, we allow for general $f_{j,i} \in \mathcal{C}^2(\mathbb{R})$.

First, we address the question of identifiability of the class of PLSEMs. Intuitively, it is evident from the respective identifiability results for linear Gaussian SEMs and CAMs that the presence of nonlinear functions in the model improves the identifiability properties of the associated DAG. We show, that this intuition is indeed correct. Our results in Chapter 3 precisely characterize how and to what extent single nonlinear additive functions improve the identifiability of the associated DAG. From that, we derive a systematic characterization of the identifiability of PLSEMs. To our knowledge, these are the first results that address and completely

characterize the identifiability of additive SEMs with Gaussian noise and non-exclusive functional type.

Second, we present an efficient score-based methodology to estimate all PLSEMs that are equivalent to a given PLSEM and derive its consistency in low- and high-dimensional settings. The ability to characterize and algorithmically learn distribution equivalence classes of PLSEMs is an important first step towards the development of a structure learning methodology for the class of PLSEMs.

Chapter 4: Estimation of causal effects in nonparametric SEMs

As a last step, we additionally drop the additivity and Gaussianity restrictions, and go back to the general SEM that we introduced in equation (1.1). As motivated in Section 1.3, the identifiability and structure learning of the associated DAG suffer from the generality of the SEM.

In Chapter 4, we ask a different question: suppose we (approximately) know the structure of the causal mechanisms (the associated DAG), is it possible to estimate causal effects under interventions for this general class of SEMs? Interestingly, this problem is much better posed than the ones of identifiability and structure learning of the associated DAG. Intuitively, the total causal effect of a single variable intervention on a response variable of interest is a one-dimensional function of the intervention value. Therefore, its estimation should not be exposed to the curse of dimensionality. We show that this is indeed the case: under suitable smoothness conditions, a specific marginal integration regression technique achieves the optimal univariate convergence rate of nonparametric function estimation for the estimation of single variable intervention effects. We propose a reasonably robust way of implementing our methodology based on an additive approximation and subsequent L_2-boosting and evaluate its performance in extensive simulation studies.

Chapter 2

Structure learning for causal additive models[1]

We develop estimation for potentially high-dimensional additive structural equation models. A key component of our approach is to decouple order search among the variables from feature or edge selection in a directed acyclic graph encoding the causal structure. We show that the former can be done with non-regularized (restricted) maximum likelihood estimation while the latter can be efficiently addressed using sparse regression techniques. Thus, we substantially simplify the problem of structure search and estimation for an important class of causal models. We establish consistency of the (restricted) maximum likelihood estimator for low- and high-dimensional scenarios, and we also allow for misspecification of the error distribution. Furthermore, we develop an efficient computational algorithm which can deal with many variables, and the new method's accuracy and

[1]This chapter is a slightly modified version of the published article Bühlmann, P., Peters, J., and Ernest, J. (2014). „CAM: Causal Additive Models, high-dimensional order search and penalized regression". *Annals of Statistics* 42 (6), pp. 2526–2556. DOI: 10.1214/14-AOS1260. Jan Ernest's main contributions (in joint work with Jonas Peters) are the conceptual development of the CAM algorithm, its implementation in the corresponding R-package CAM and the realization of the numerical experiments. The theoretical results in Section 2.4 have been derived and proved by Peter Bühlmann and major parts of the paper were written by the two co-authors. To motivate the algorithm, we include the theoretical results (as given in the main text of Bühlmann et al. (2014)), but omit their proofs. All proofs can be found in the supplement to the original article.

performance is illustrated on simulated and real data.

2.1 Introduction

Inferring causal relations and effects is an ambitious but important task in virtually all areas of science. In absence of prior information about underlying structure, the problem is plagued, among other things, by identifiability issues (cf. Pearl, 2000; Spirtes et al., 2000) and the sheer size of the space of possible models, growing super-exponentially in the number of variables, leading to major challenges with respect to computation and statistical accuracy. Our approach is generic, taking advantage of the tools in sparse regression techniques (cf. Bühlmann and van de Geer, 2011; Hastie et al., 2009) which have been successively established in recent years.

More precisely, we consider p random variables $X_1, \ldots, X_p$ whose distribution is Markov with respect to an underlying causal directed acyclic graph (causal DAG). We assume that all variables are observed, that is, there are no hidden variables, and that the causal influence diagram doesn't allow for directed cycles. Generalizations to include hidden variables, for example, unobserved confounders, or directed cycles are briefly discussed in Section 2.8.2. To formalize a model, one can use the concepts of graphical modeling (cf. Lauritzen, 1996) or structural equation models (cf. Pearl, 2000). The approaches are equivalent in the nonparametric or multivariate Gaussian case, but this is not true anymore when placing additional restrictions which can be very useful (cf. Peters and Bühlmann, 2014; Peters et al., 2014; Shimizu et al., 2006). We use here the framework of structural equation models.

2.1.1 Problem and main idea

Our goal is estimation and structure learning for structural equation models, or of the corresponding Markov equivalence class of an underlying DAG. In particular, we focus on causal additive models, that is, the structural equations are additive in the variables and error terms. The model has the nice property that the underlying structure and the corresponding parameters are identifiable from the observational distribution. Furthermore, we can view it as an extension of linear Gaussian structural equation models by allowing for nonlinear additive functions.

In general, the problem of structure learning (and estimation of corresponding parameters) can be addressed by a variety of algorithms and methods: in the frequentist setting, the most widely used procedures for structure learning (and corresponding parameters) are greedy equivalence search for computing the BIC-regularized maximum likelihood estimator (Chickering, 2002) or the PC-algorithm using multiple conditional independence testing (Spirtes et al., 2000). However, for the latter, the constraint of additive structural equations cannot be (easily) respected, and regarding the former, maximum likelihood estimation among all (e.g., linear Gaussian) DAG models is computationally challenging and statistical guarantees for high-dimensional cases (and for uniform convergence with respect to a class of distributions) are only available under rather strong assumptions (van de Geer and Bühlmann, 2013).

Our proposed approach for estimation and selection of additive structural equation models is based on the following simple idea which is briefly mentioned and discussed in Teyssier and Koller (2005) and Schmidt et al. (2007). If the order among the variables would be known, the problem boils down to variable selection in multivariate (potentially nonlinear) regression, see formula (2.5). The latter is very well understood: for example, we can follow the route of hypothesis testing in additive models, or sparse regression can be used for additive models (Meier et al., 2009; Ravikumar et al., 2009; Yuan and Lin, 2006). Thus, the only remaining task is to estimate the order among the variables. We show here that this can be done via the maximum likelihood principle, and we establish its consistency. In particular, for low or "mid"-dimensional problems, there is no need to consider a penalized likelihood approach. The same holds true for high-dimensional settings when using a preliminary neighborhood selection and then employing a corresponding restricted maximum likelihood estimator. Therefore, we can entirely decouple the issue of order estimation without regularization and variable selection in sparse regression with appropriate regularization. This makes our approach very generic, at least within the framework where the underlying DAG and a corresponding order of the variables are identifiable from the joint distribution. Empirical results in Section 2.6 support that we can do much more accurate estimation than for non-identifiable models such as the popular linear Gaussian structural equation model. On the superficial level, our approach can be summarized as follows:

1. Mainly for high-dimensional settings: preliminary neighborhood selection for estimating a superset of the skeleton of the underlying

DAG. This is done by additive regression of one variable against all others. See Section 2.3.1.

2. Order search for the variables (or best permutation for the indices of the variables) using (restricted) maximum likelihood estimation based on an additive structural equation model with Gaussian errors: the restricted version is employed if the preliminary neighborhood selection in Step 1 is used, and the order search is then restricted to the structure of the superset of the skeleton. See Sections 2.2.4 and 2.3.2.

3. Based on the estimated order of the variables in Step 2, sparse additive regression is used for estimating the functions in an additive structural equation model. See Section 2.2.5.

2.1.2 Related work

We consider (nonlinear) additive structural equation models. As natural extensions of linear structural equation models, they are attractive for many applications, see Imoto et al. (2002). Identifiability results for this model class have been recently derived (Mooij et al., 2009; Peters et al., 2014). The approach in Mooij et al. (2009) is based on conditional independence testing and is limited to small dimensions with a few variables only. Instead of multiple testing of conditional independences, we propose and develop maximum likelihood estimation in a semiparametric additive structural equation model with Gaussian noise variables: fitting such a model is often appropriate in situations where the sample size is not too large, and we present here for the first time the practical feasibility of fitting additive models in the presence of many variables. An extension of our additive structural equation model with Gaussian errors to the case with a nonparametric specification of the error distribution is presented in Nowzohour and Bühlmann (2016), but the corresponding maximum likelihood estimator is analyzed (and feasible) for problems with a small number of variables only. When the order of the variables is known, which is a much simpler and different problem than what we consider here, Voorman et al. (2014) provide consistency results for additive structural equation models.

A key aspect of our method is that we decouple regularization for feature selection and order estimation with non-regularized (restricted) maximum likelihood. The former is a well-understood subject thanks to the broad literature in sparse regression and related techniques (cf. Meinshausen and

Bühlmann, 2006; Tibshirani, 1996; Wainwright, 2009; Yuan and Lin, 2006; Zhao and Yu, 2006; Zou, 2006). Regarding the latter issue about order selection, a recent analysis in van de Geer (2014) extends our low-dimensional consistency result for the (non-restricted) maximum likelihood estimator to the scenario where the number of variables can grow with sample size, in the best case essentially as fast as $p = p(n) = o(n)$. The treatment of the high-dimensional case with a restricted maximum likelihood approach is new here, and we also present the first algorithm and empirical results for fitting low- and high-dimensional causal additive models (CAMs).

2.2 Additive structural equation models

Consider the general structural equation model (SEM):

$$X_j = f_j(X_{\mathrm{pa}_D(j)}, \varepsilon_j),$$

where $\mathrm{pa}_D(j)$ denotes the set of parents of node j in DAG D, f_j is a function from $\mathbb{R}^{|\mathrm{pa}_D(j)|+1} \to \mathbb{R}$ and $\varepsilon_1, \ldots, \varepsilon_p$ are (random) noise variables which are assumed to be (mutually) independent. Thus, a SEM is specified by an underlying (causal) structure in terms of a DAG D, the functions $f_j(\cdot)$ $(j = 1, \ldots, p)$ and the distributions of ε_j $(j = 1, \ldots, p)$. Most parts of this chapter can be interpreted in absence of causal inference issues: clearly though, the main motivations are understanding models and developing novel procedures allowing for causal or interventional statements, and if we do so, we always assume that the structural equations remain unchanged under interventions at one or several variables (cf. Pearl, 2000). The model above is often too general, due to problems of identifiability and the difficulty of estimation (curse of dimensionality) of functions in several variables.

Our main focus is on a special (and more practical) case of the model above, namely the additive SEM with potentially misspecified Gaussian errors:

$$X_j = \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + \varepsilon_j, \tag{2.1}$$

where $\varepsilon_1, \ldots, \varepsilon_p$ are independent with $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \sigma_j^2 > 0, (j = 1, ..., p)$, and $f_{j,k}(\cdot)$ are smooth functions from $\mathbb{R} \to \mathbb{R}$ with $\mathbb{E}[f_{j,k}(X_k)] = 0$ for

all j, k. A special case thereof is the linear Gaussian SEM

$$X_j = \sum_{k \in \mathrm{pa}_D(j)} \beta_{j,k} X_k + \varepsilon_j, \tag{2.2}$$

with $\varepsilon_1, \ldots, \varepsilon_p$ independent with $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $\sigma_j^2 > 0$, $(j = 1, \ldots, p)$. Although model (2.2) is a special case of (2.1), there are interesting differences with respect to identifiability. If all functions $f_{j,k}(\cdot)$ are nonlinear, the DAG is identifiable from the distribution P of $X_1, \ldots, X_p$ (Peters et al., 2014, Corollary 31). We explicitly state this result as a lemma since we will make use of it later on.

Lemma 1 (Corollary 31 in Peters et al. (2014)[2])**.** *Consider a distribution P that is generated by model* (2.1) *with DAG D and nonlinear, three times differentiable functions $f_{j,k}$. Then, any distribution Q that is generated by* (2.1) *with a different DAG $D' \neq D$ and non-constant, three times differentiable functions $f'_{j,k}$ is different from P: we have $Q \neq P$.*

This result does not hold, however, for a general SEM or for a linear Gaussian SEM as in (2.2); one can then only identify the Markov equivalence class of the DAG D^0, assuming faithfulness. An exception arises when assuming same error variances $\sigma_j^2 \equiv \sigma^2$ for all j in (2.2) which again implies identifiability of the DAG D^0 from P (Peters and Bühlmann, 2014). In the sequel, we consider the fully identifiable case of model (2.1).

2.2.1 The likelihood

We slightly re-write model (2.1) as

$$\begin{aligned}
&X_j = \sum_{k \in \mathrm{pa}_D(j)} f_{j,k}(X_k) + \varepsilon_j = \sum_{k \neq j} f_{j,k}(X_k) + \varepsilon_j \quad (j = 1, \ldots, p), \\
&f_{j,k}(\cdot) \not\equiv 0 \text{ if and only if there is a directed edge } k \to j \text{ in } D, \\
&\mathbb{E}[f_{j,k}(X_k)] = 0 \text{ for all } j, k, \\
&\varepsilon_1, \ldots, \varepsilon_p \text{ independent and } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2),\ \sigma_j^2 > 0.
\end{aligned} \tag{2.3}$$

The structure of the model, or the so-called active set, $\{(j, k);\ f_{j,k} \not\equiv 0\}$ is identifiable from the distribution P (Peters et al., 2014, Corollary 31).

[2]Corollary 31 in Peters et al. (2014) contains a slightly different statement using "nonlinear" instead of "non-constant". The proof, however, stays exactly the same.

Denote by θ the infinite-dimensional parameter with additive functions and error variances, that is:

$$\theta = (f_{1,2}, \ldots, f_{1,p}, f_{2,1}, \ldots, f_{p,p-1}, \sigma_1, \ldots, \sigma_p).$$

Moreover, we denote by D^0 the true DAG and by θ^0 (and $\{f^0_{j,k}\}$, $\{\sigma^0_j\}$) the true infinite-dimensional parameter corresponding to the data-generating true distribution. We use this notation whenever it is appropriate to make statements about the true underlying DAG or parameter.

The density $p_\theta(\cdot)$ for the model (2.3) is of the form:

$$\log(p_\theta(x)) = \sum_{j=1}^{p} \log\left(\frac{1}{\sigma_j}\varphi\left(\frac{x_j - \sum_{k\neq j} f_{j,k}(x_k)}{\sigma_j}\right)\right),$$

where $\varphi(\cdot)$ is the density of a standard Normal distribution. Furthermore,

$$\sigma_j^2 = \mathbb{E}[(X_j - \sum_{k\neq j} f_{j,k}(X_k))^2],$$

and the expected negative log-likelihood is:

$$\mathbb{E}_\theta[-\log p_\theta(X)] = \sum_{j=1}^{p} \log(\sigma_j) + C, \quad C = \frac{p}{2}\log(2\pi) + \frac{p}{2}.$$

2.2.2 The function class

We assume that the functions in model (2.1) or (2.3) are from a class of smooth functions: $\mathcal{F}$ is a subset of $L_2(P_j)$, where P_j is the marginal distribution for any $j = 1, \ldots, p$; assume that it is closed with respect to the $L_2(P_j)$ norm. Furthermore,

$$\mathcal{F} \subseteq \{f : \mathbb{R} \to \mathbb{R},\ f \in C^\alpha,\ \mathbb{E}[f(X)] = 0\},$$

where C^α denotes the space of α-times differentiable functions and the random variable X is a placeholder for the variables X_j, $j = 1, \ldots, p$. Note that this is a slight abuse of notation since $\mathcal{F}$ does not specify the variable X; it becomes clear from the context.

Consider also basis functions $\{b_r(\cdot);\ r = 1, \ldots, a_n\}$ with $a_n \to \infty$ sufficiently slowly, for example, B-splines or regression splines. Consider fur-

ther the space

$$\mathcal{F}_n = \left\{ f \in \mathcal{F},\ f = c + \sum_{r=1}^{a_n} \alpha_r b_r(\cdot) \text{ with } c, \alpha_r \in \mathbb{R}\,(r = 1, \ldots, a_n) \right\}. \tag{2.4}$$

We allow for constants c to enforce mean zero for the whole function. Furthermore, the basis functions can be the same for all variables X_j, $j = 1, \ldots, p$. For theoretical analysis, we assume that $\mathcal{F}_n$ is deterministic and does not depend on the data. Then, $\mathcal{F}_n$ is closed. Furthermore, the space of additive functions is denoted by

$$\mathcal{F}^{\oplus \ell} = \left\{ f : \mathbb{R}^\ell \to \mathbb{R};\ f(x) = \sum_{k=1}^{\ell} f_k(x_k),\ f_k \in \mathcal{F} \right\},$$

$$\mathcal{F}_n^{\oplus \ell} = \left\{ f : \mathbb{R}^\ell \to \mathbb{R};\ f(x) = \sum_{k=1}^{\ell} f_k(x_k),\ f_k \in \mathcal{F}_n \right\},$$

where $\ell = 2, \ldots, p$. Clearly $\mathcal{F}_n^{\oplus \ell} \subseteq \mathcal{F}^{\oplus \ell}$. In our definitions, we assume that the functions in $\mathcal{F}$ and $\mathcal{F}_n$ have expectation zero. Of course, this depends on the variables in the arguments of the functions. For example, when requiring $\mathbb{E}[f(X_j)] = 0$ for $f \in \mathcal{F}$, the function class $\mathcal{F} = \mathcal{F}_j$ depends on the index j due to the mean zero requirement; and likewise $\mathcal{F}^{\oplus \ell}$ depends on the indices of the variables occurring in the corresponding additive function terms. We drop this additional dependence on the index of variables as it does not cause any problems in methodology or theory.

Later, we consider projections of distributions onto the spaces $\mathcal{F}^{\oplus \ell}$ and $\mathcal{F}_n^{\oplus \ell}$, see (2.6). We assume throughout the chapter that these spaces are closed with respect to the L_2 norm. Lemma 2 guarantees this condition by requiring an analogue of a minimal eigenvalue assumption.

Lemma 2. *Let the distribution P be generated according to* (2.1) *and assume that there is a $\phi^2 > 0$ such that for all $\gamma \in \mathbb{R}^p$*

$$\Big\| \sum_{j=1}^{p} \gamma_j f_j(X_j) \Big\|_{L_2}^2 \geq \phi^2 \, \|\gamma\|^2 \quad \text{for all } f_j \in \mathcal{F} \text{ with } \|f_j(X_j)\|_{L_2} = 1\,.$$

For any subset $I \subseteq \{1, \ldots, p\}$ of ℓ variables the spaces $\mathcal{F}^{\oplus \ell}$ and $\mathcal{F}_n^{\oplus \ell}$ are then closed with respect to the $L_2(P_I)$ norm. Here, P_I denotes the marginal distribution over all variables in I.

The question of closedness of additive models has also been studied in Breiman and Friedman (1985), for example; see also Rényi (1959).

2.2.3 Order of variables and the likelihood

We can permute the variables, inducing a different ordering; in the sequel, we use both terminologies, permutations and order search, which mean the same thing. For a permutation π on $\{1, \ldots, p\}$, define:

$$X^{\pi}, \quad X_j^{\pi} = X_{\pi(j)}.$$

There is a canonical correspondence between permutations and fully connected DAGs: For any permutation π we can construct a DAG D^{π}, in which each variable $\pi(k)$ has a directed arrow to all $\pi(j)$ with $j > k$. The node $\pi(1)$ has no parents and is called the source node. For a given DAG D^0 we define the set of true permutations as

$$\Pi^0 = \{\pi^0;\ \text{the fully connected DAG } D^{\pi^0} \text{ is a super-DAG of } D^0\},$$

where a super-DAG of D^0 is a DAG whose set of directed edges is a superset of the one corresponding to D^0. If the true DAG D^0 is not fully connected, there is typically more than one true order or permutation, that is the true order is typically not unique. It is apparent that any true ordering or permutation π^0 allows for a lower-triangular (or autoregressive) representation of the model in (2.3):

$$X_j^{\pi^0} = \sum_{k=1}^{j-1} f_{j,k}^{\pi^0}(X_k^{\pi^0}) + \varepsilon_j^{\pi^0} \ \ (j = 1, \ldots, p), \tag{2.5}$$

where $f_{j,k}^{\pi^0}(\cdot) = f_{\pi^0(j),\pi^0(k)}^0(\cdot)$ and $\varepsilon_j^{\pi^0} = \varepsilon_{\pi^0(j)}^0$, that is, with permuted indices in terms of the original quantities in (2.3). If all functions $f_{j,k}(\cdot)$ are nonlinear, the set of true permutations is identifiable from the distribution (Peters et al., 2014, Corollary 33), and Π^0 consists of all orderings of the variables which allow for a lower-triangular representation (2.5). We will exploit this fact in order to provide a consistent estimator $\hat{\pi}_n$ of the ordering: under suitable assumptions the probability that $\hat{\pi}_n \in \Pi^0$ converges to one.

Remark 1. *For the linear Gaussian SEM* (2.2)*, all orderings allow for a lower-triangular representation* (2.5)*, even those that are not in* Π^0*. Thus, we cannot construct a consistent estimator in the above sense. However, assuming faithfulness of the true distribution, the orderings of variables which are consistent with the arrow directions in a DAG of the Markov equivalence class of the true DAG* D^0 *lead to sparsest representations with fewest number of non-zero coefficients.*

In principle, one can check whether the data come from a linear Gaussian SEM. Lemma 1 guarantees that if this is case, there is no CAM with nonlinear functions yielding the same distribution. Thus, if the structural equations of the estimated DAG look linear with Gaussian noise, one could decide to output the Markov equivalence class instead of the DAG. One would need to quantify closeness to linearity and Gaussianity with, for example, a test: this would be important for practical applications, but its precise implementation lies beyond the scope of this work.

In the sequel, it is helpful to consider the true underlying parameter θ^0 with corresponding nonlinear function $f^0_{j,k}$ and error variances $(\sigma^0_j)^2$. For any permutation $\pi \notin \Pi^0$, we consider the projected parameters, defined as

$$\theta^{\pi,0} = \operatorname*{argmin}_{\theta^\pi} \mathbb{E}_{\theta^0}[-\log(p^\pi_{\theta^\pi}(X))],$$

where the density $p^\pi_{\theta^\pi}$ is of the form:

$$\log(p^\pi_{\theta^\pi}(x)) = \log(p_{\theta^\pi}(x^\pi)) = \sum_{j=1}^{p} \log\left(\frac{1}{\sigma^\pi_j}\varphi\left(\frac{x^\pi_j - \sum_{k=1}^{j-1} f^\pi_{j,k}(x^\pi_k)}{\sigma^\pi_j}\right)\right).$$

(Note that if $\pi \in \Pi^0$, then $\theta^{\pi,0} = \theta^0$.) For such a misspecified model with wrong order $\pi \notin \Pi^0$, we have

$$\begin{aligned}\{f^{\pi,0}_{j,k}\}_{k=1,\ldots,j-1} &= \operatorname*{argmin}_{g_{j,k}\in\mathcal{F},\ k=1,\ldots,j-1} \mathbb{E}_{\theta^0}\left[\left(X^\pi_j - \sum_{k=1}^{j-1} g_{j,k}(X^\pi_k)\right)^2\right] \qquad (2.6)\\ &= \operatorname*{argmin}_{g_j\in\mathcal{F}\oplus j-1} \mathbb{E}_{\theta^0}[(X^\pi_j - g_j(X^\pi_1,\ldots,X^\pi_{j-1}))^2].\end{aligned}$$

It holds that:

$$\begin{aligned}(\sigma^{\pi,0}_j)^2 &= \operatorname*{argmin}_{\sigma^2}\left(\log(\sigma) + \frac{1}{2\sigma^2}\mathbb{E}_{\theta^0}\left[\left(X^\pi_j - \sum_{k=1}^{j-1} f^{\pi,0}_{j,k}(X^\pi_k)\right)^2\right]\right)\\ &= \mathbb{E}_{\theta^0}\left[\left(X^\pi_j - \sum_{k=1}^{j-1} f^{\pi,0}_{j,k}(X^\pi_k)\right)^2\right]. \qquad (2.7)\end{aligned}$$

The two displayed formulae above show that autoregression with the wrong order π leads to the projected parameters $\{f^{\pi,0}_{j,k}\}$ and $\{(\sigma^{\pi,0}_j)^2\}$. Finally,

we obtain:

$$\mathbb{E}_{\theta^0}[-\log(p^{\pi}_{\theta^{\pi},0}(X))] = \sum_{j=1}^{p} \log(\sigma_j^{\pi,0}) + C, \quad C = \frac{p}{2}\log(2\pi) + \frac{p}{2}.$$

All true permutations $\pi \in \Pi^0$ correspond to super-DAGs of the true DAG and therefore, all of them lead to the minimal expected log-likelihood $\mathbb{E}_{\theta^0}[-\log(p^{\pi}_{\theta^{\pi},0}(X))] = \mathbb{E}_{\theta^0}[-\log(p_{\theta^0}(X))]$. The permutations $\pi \notin \Pi^0$, however, cannot lead to a smaller expected negative log-likelihood (since it would lead to a negative KL-divergence between the true and best projected distribution). Let us therefore define

$$\xi_p := \min_{\pi \notin \Pi^0} p^{-1}(\mathbb{E}_{\theta^0}[-\log(p^{\pi}_{\theta^{\pi},0}(X))] - \mathbb{E}_{\theta^0}[-\log(p_{\theta^0}(X))]) \geq 0. \tag{2.8}$$

If all true functions $f^0_{j,k}$ are nonlinear, we obtain $\xi_p > 0$ as follows.

Lemma 3. *Consider a distribution P that allows for a density p with respect to the Lebesgue measure and is generated by model* (2.1) *with DAG D^0 and nonlinear, three times differentiable functions $f^0_{j,k}$. Assume further the condition from Lemma 2. Then $\xi_p > 0$.*

Proof. Because of the closedness of $\mathcal{F}^{\oplus j}$ (Lemma 2), the minimum in (2.6) is obtained for some functions $f_{j,k}$. Without loss of generality, we can assume that all constant additive components are zero. But then $\xi_p = 0$ would contradict Lemma 1. □

The number ξ_p describes the degree of separation between the true model and misspecification when using a wrong permutation. As discussed in Remark 1, $\xi_p = 0$ for the case of linear Gaussian SEMs. Formula (2.8) can be expressed as

$$\xi_p = \min_{\pi \notin \Pi^0} p^{-1} \sum_{j=1}^{p} (\log(\sigma_j^{\pi,0}) - \log(\sigma_j^0)) \geq 0. \tag{2.9}$$

Remark 2. *Especially for situations where p is very large so that the factor p^{-1} is small, requiring a lower bound $\xi_p > 0$ can be overly restrictive. Instead of requiring a gap with the factor p^{-1} between the likelihood scores of the true distribution and all distributions corresponding to permutations, one can weaken this as follows. Let $H(D, D^0) = \{j;\ \mathrm{pa}_{D^0}(j) \not\subseteq \mathrm{pa}_D(j)\}$. We require that*

$$\xi_p' := \min_{D \neq D^0} |H(D, D^0)|^{-1} \sum_{j \in H(D,D^0)} (\log(\sigma_j^{D,0}) - \log(\sigma_j^0)) \geq 0, \tag{2.10}$$

where $(\sigma_j^{D,0})^2$ *is the error variance in the best additive approximation of* X_j *based on* $\{X_k;\ k \in \mathrm{pa}_D(j)\}$. *Such a weaker gap condition is proposed in Loh and Bühlmann (2014, Section 5.2). All our theoretical results still hold when replacing statements involving* ξ_p *in* (2.9) *by the corresponding statements with* ξ_p' *in* (2.10).

2.2.4 Maximum likelihood estimation for order: Low-dimensional setting

We assume having n i.i.d. realizations $X^{(1)}, \ldots, X^{(n)}$ from model (2.3). For a $n \times 1$ vector $x = (x^{(1)}, \ldots, x^{(n)})^T$, we denote by $\|x\|_{(n)}^2 = n^{-1} \sum_{i=1}^n (x^{(i)})^2$. Depending on the context, we sometimes denote by $\hat{f}$ a function and sometimes an $n \times 1$ vector evaluated at (the components of) the data points $X^{(1)}, \ldots, X^{(n)}$; and similarly for X_j^π. We consider the unpenalized maximum likelihood estimator:

$$\hat{f}_j^\pi = \operatorname*{argmin}_{g_j \in \mathcal{F}_n^{\oplus j-1}} \left\| X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi) \right\|_{(n)}^2,$$

$$(\hat{\sigma}_j^\pi)^2 = \left\| X_j^\pi - \sum_{k=1}^{j-1} \hat{f}_{j,k}^\pi(X_k^\pi) \right\|_{(n)}^2.$$

Denote by $\hat{\pi}$ a permutation which minimizes the unpenalized negative log-likelihood:

$$\hat{\pi} \in \operatorname*{argmin}_\pi \sum_{j=1}^p \log(\hat{\sigma}_j^\pi). \tag{2.11}$$

The estimation of $\hat{f}_j^\pi$ is based on $\mathcal{F}_n$ with pre-specified basis functions $b_r(\cdot)$ with $r = 1, \ldots, a_n$. In practice, the basis functions could depend on the predictor variable or on the order of variables, for example, when choosing the knots in regression splines. The classical choice for the number of basis functions is $a_n \asymp n^{1/5}$ for twice differentiable functions: here, and as explained in Section 2.4, however, a smaller number such as $a_n = O(1)$ to detect some nonlinearity might be sufficient for estimation of the true underlying order.

2.2.5 Sparse regression for feature selection

Section 2.4 presents assumptions and results ensuring that with high probability $\hat{\pi} = \pi^0$ for some $\pi^0 \in \Pi^0$. With such an estimated order $\hat{\pi}$, we obtain a complete super-DAG $D^{\hat{\pi}}$ of the underlying DAG D^0 in (2.3), where the parents of a node $\hat{\pi}(j)$ are defined as $\mathrm{pa}_{D^{\hat{\pi}}}(\hat{\pi}(j)) = \{\hat{\pi}(k);\ k < j\}$ for all j. We can pursue consistent estimation of intervention distributions based on $D^{\hat{\pi}}$ without any additional need to find the true underlying DAG D^0; see Section 2.2.6.

However, we can improve statistical efficiency for estimating the intervention distribution when it is ideally based on the true DAG D^0 or realistically a not too large super-DAG $\hat{D}^{\hat{\pi}} \supseteq D^0$. The task of estimating such a super-DAG $\hat{D}^{\hat{\pi}} \supseteq D^0$ is conceptually straightforward: starting from the complete super-DAG $D^{\hat{\pi}}$ of D^0 as discussed above, we can use model selection or a penalized multivariate (auto-) regression technique in the model representation (2.5). For additive model fitting, we can either use hypothesis testing for additive models (Marra and Wood, 2011) or the Group Lasso (Ravikumar et al., 2009), or its improved version with a sparsity-smoothness penalty proposed in Meier et al. (2009). All the techniques mentioned above perform variable selection, where we denote by

$$\hat{D}^{\hat{\pi}} = \{(\hat{\pi}(k), \hat{\pi}(j));\ \hat{f}^{\hat{\pi}}_{j,k} \not\equiv c\},$$

(the constant $c = 0$ when assuming that $\hat{f}^{\hat{\pi}}_{j,k}$ have mean zero when evaluated over all data-points) the selected variables indexed in the original order (we obtain estimates $\hat{f}^{\hat{\pi}}_{j,k}$ in the representation (2.5) with correspondence to the indices $\hat{\pi}(k), \hat{\pi}(j)$ in the original order); we identify these selected variables in $\hat{D}^{\hat{\pi}}$ as the edge set of a DAG. For example with the Group Lasso, assuming some condition avoiding near collinearity of functions, that is, a compatibility condition for the Group Lasso (Bühlmann and van de Geer, 2011, Chapter 5.6, Theorem 8.2), and that the ℓ_2-norms of the non-zero functions are sufficiently large, we obtain the screening property (since we implicitly assume that $\hat{\pi} \in \Pi^0$ with high probability): with high probability and asymptotically tending to one,

$$\hat{D}^{\hat{\pi}} \supseteq D^0 = \{(k, j);\ f^0_{j,k} \not\equiv 0\} \tag{2.12}$$

saying that all relevant variables (i.e., edges) are selected. Similarly with hypotheses testing, assuming that the non-zero $f^0_{j,k}$ have sufficiently large ℓ_2-norms, we also obtain that (2.12) holds with high probability.

The same argumentation applies if we use $D^{\hat{\pi}}_{\text{restr}}$ from Section 2.3.2 instead of $D^{\hat{\pi}}$ as an initial estimate. This then results in $\hat{D}^{\hat{\pi}}_{\text{restr}}$, replacing $\hat{D}^{\hat{\pi}}$ above.

2.2.6 Consistent estimation of causal effects

The property in (2.12) has an important implication for causal inference[3]: all estimated causal effects and estimated intervention distributions based on the estimated DAG $\hat{D}^{\hat{\pi}}$ are consistent. In fact, using the do-calculus (cf. Pearl, 2000, (3.10)), we have for the single intervention (at variable X_k) distribution for X_j, for all $j \neq k$:

$$p_{D^0}(x_j|\text{do}(X_k = x)) = p_{\hat{D}^{\hat{\pi}}}(x_j|\text{do}(X_k = x)), \text{ for all } x,$$

where $p_D(\cdot|\text{do}(\cdot))$ denotes the intervention density based on a DAG D.

We note that the screening property (2.12) also holds when replacing $\hat{D}^{\hat{\pi}}$ with the full DAG induced by $\hat{\pi}$, denoted by $D^{\hat{\pi}}$. Thus, the feature selection step in Section 2.2.5 is not needed to achieve consistent estimation of causal effects. However, a smaller DAG $D^0 \subseteq \hat{D}^{\hat{\pi}} \subseteq D^{\hat{\pi}}$ typically leads to better (more statistically efficient) estimates of the interventional distributions than the full DAG $D^{\hat{\pi}}$.

2.3 Restricted maximum likelihood estimation: computational and statistical benefits

We present here maximum likelihood estimation where we restrict the permutations, instead of searching over all permutations in (2.11). Such a restriction makes the computation more tractable, and it is also statistically crucial when dealing with high-dimensional settings where $p > n$.

2.3.1 Preliminary neighborhood selection

We first perform neighborhood selection with additive models, following the general idea in Meinshausen and Bühlmann (2006) for the linear Gaus-

[3]We assume that interventions at variables do not change the other structural equations, and that there are no unobserved hidden (e.g., confounder) variables.

sian case. We pursue variable selection in an additive model of X_j versus all other variables $X_{\{-j\}} = \{X_k;\ k \neq j\}$: a natural method for such a feature selection is the Group Lasso for additive models (Ravikumar et al., 2009), ideally with a sparsity-smoothness penalty (Meier et al., 2009); see also Voorman et al. (2014). This provides us with a set of variables

$$\hat{A}_j \subseteq \{1, \ldots, p\} \setminus j,$$

which denotes the selected variables in the estimated conditional expectation

$$\hat{\mathbb{E}}_{\text{add}}[X_j | X_{\{-j\}}] = \sum_{k \in \hat{A}_j} \hat{h}_{jk}(X_k)$$

with functions $\hat{h}_{jk}$ satisfying $n^{-1} \sum_{i=1}^n \hat{h}_{jk}(X_k^{(i)}) = 0$ (i.e., a possible intercept is subtracted already): that is,

$$\hat{A}_j = \{k;\ k \neq j,\ \hat{h}_{j,k} \not\equiv 0\}.$$

We emphasize that the functions $\hat{h}_{j,k}(\cdot)$ are different from $\hat{f}^{\pi}_{j,k}(\cdot)$ in Section 2.2.4 because for the former, the additive regression is against all other variables.

We give conditions in Section 2.4.2, see Lemma 4, ensuring that the neighborhood selection set contains the parental variables from the structural equation model in (2.1) or (2.3), that is, $\hat{A}_j \supseteq \text{pa}(j)$.

2.3.2 Restricted maximum likelihood estimator

We restrict the space of permutations in the definition of (2.11) such that the permutations are "compatible" with the neighborhood selection sets $\hat{A}_j$. Note that for the estimator $\hat{\sigma}^{\pi}_j$ in (2.11), we regress $X_{\pi(j)}$ against $\{X_k;\ k \in \{\pi(j-1), \ldots, \pi(1)\}\}$. We restrict here the set of regressors to the indices $R_{\pi,j} = \{\pi(j-1), \ldots, \pi(1)\} \cap \hat{A}_{\pi(j)}$ and calculate the $\pi(j)$-th term of the log-likelihood using this set of regressors $X_{R_{\pi,j}} = \{X_k;\ k \in R_{\pi,j}\}$. More precisely, we estimate

$$\hat{f}^{\pi,R}_j = \operatorname*{argmin}_{g_{j,k} \in \mathcal{F}_n} \left\| X^{\pi}_j - \sum_{k;\, \pi(k) \in R_{\pi,j}} g_{j,k}(X^{\pi}_k) \right\|^2_{(n)},$$

$$(\hat{\sigma}^{\pi,R}_j)^2 = \left\| X^{\pi}_j - \sum_{k;\, \pi(k) \in R_{\pi,j}} \hat{f}^{\pi,R}_{j,k}(X^{\pi}_k) \right\|^2_{(n)},$$

and the restricted maximum likelihood estimator is

$$\hat{\pi} \in \operatorname*{argmin}_{\pi} \sum_{j=1}^{p} \log(\hat{\sigma}_j^{\pi,R}). \tag{2.13}$$

If $\max_j |\hat{A}_j| < n$, the estimators $\hat{\sigma}_j^{\pi,R}$ are well-defined.

The computation of the restricted maximum likelihood estimator in (2.13) is substantially easier than for the unrestricted MLE (2.11) if $\max_j |\hat{A}_j|$ is small (which is ensured if the true neighborhoods are sparse). The set of all permutations can be partitioned in equivalence classes $\cup_r \mathcal{R}_r$ and the minimization in (2.13) can be restricted to single representatives of each equivalence class $\mathcal{R}_r$. The equivalence relation can be formulated with a restricted DAG D^{π}_{restr} whose parental set for node $\pi(j)$ equals $\text{pa}_{D^{\pi}_{\text{restr}}}(\pi(j)) = R_{\pi,j}$. We then have that

$$\pi \sim \pi' \text{ if and only if } D^{\pi}_{\text{restr}} = D^{\pi'}_{\text{restr}}.$$

Computational details are described in Section 2.5.

2.4 Consistency in correct and misspecified models

We prove consistency for the ordering among variables in additive structural equation models, and under an additional identifiability assumption even for the case where the model is misspecified with respect to the error distribution or when using highly biased function estimation.

2.4.1 Unrestricted MLE for low-dimensional settings

We first consider the low-dimensional setting where $p < \infty$ is fixed and $n \to \infty$, and we establish consistency of the unrestricted MLE in (2.11). We assume the following:

(A1) Consider a partition of the real line

$$\mathbb{R} = \cup_{m=1}^{\infty} I_m$$

in disjoint intervals I_m. The individual functions in $\mathcal{F}$ are α-times differentiable, with $\alpha \geq 1$, whose derivatives up to order α are bounded in absolute value by M_m in I_m.

(A2) Tail and moment conditions:

(i) For $V = 1/\alpha$ and M_m as in (A1):

$$\sum_{m=1}^{\infty} \left(M_m^2 \mathbb{P}[X_j \in I_m]\right)^{V/(V+2)} < \infty, \ j = 1, \ldots, p.$$

(ii)

$$\mathbb{E}|X_j|^4 < \infty, \ j = 1, \ldots, p$$
$$\sup_{f \in \mathcal{F}} \mathbb{E}|f(X_j)|^4 < \infty, \ j = 1, \ldots, p.$$

(A3) The error variances satisfy $(\sigma_j^{\pi,0})^2 > 0$ for all $j = 1, \ldots, p$ and all π.

(A4) The true functions $f_{j,k}^0$ can be approximated on any compact set $\mathcal{C} \subset \mathbb{R}$: for all $k \in \mathrm{pa}_{D^0}(j)$, $j = 1, \ldots, p$,

$$\mathbb{E}[(f_{j,k}^0(X_k) - f_{n;j,k}^0(X_k))^2 I(X_k \in \mathcal{C})] = o(1),$$

where

$$f_{n;j}^0 = \underset{g_j \in \mathcal{F}_n^{\oplus j-1}}{\operatorname{argmin}} \ \mathbb{E}\left[\left(X_j - \sum_{k \in \mathrm{pa}_{D^0}(j)} g_{j,k}(X_k)\right)^2\right].$$

All assumptions are not very restrictive. The second part of assumption (A2)(ii) holds if we assume, for example, a bounded function class $\mathcal{F}$, or if $|f(x)| \asymp |x|$ as $|x| \to \infty$ for all $f \in \mathcal{F}$.

Theorem 1. *Consider an additive structural equation model as in* (2.3). *Assume (A1)-(A4) and* $\xi_p > 0$ *in* (2.8) *(see also Lemma 3 and Remark 2). Then we have*

$$\mathbb{P}[\hat{\pi} \in \Pi^0] \to 1 \qquad (n \to \infty).$$

A proof is given in the supplement to Bühlmann et al. (2014). Theorem 1 says that one can find a correct order among the variables without pursuing feature or edge selection for the structure in the SEM.

Remark 3. *Studying near non-identifiable models, for example, near linearity in a Gaussian structural equation model, can be modeled by allowing* $\xi_p = \xi_{n,p}$ *to converge to zero as* $n \to \infty$. *If one requires* $\xi_{n,p} \gg n^{-1/2}$, *the*

statement of Theorem 1 still holds. We note that Theorem 3 for the high-dimensional case implicitly allows $\xi_p = \xi_{p_n}$ to change with sample size n. However, it is a non-trivial issue to translate such a condition in terms of closeness of one or several nonlinear functions $f^0_{j,k}$ to their closest linear approximations. Similarly, if some error variances $\sigma^{\pi,0}_j$ would be close to zero (e.g., converge to zero as $n \to \infty$ asymptotically), this could cause identifiability problems such that ξ_p might be close to (e.g., converge fast to) zero.

Related to Remark 3 is the question about uniform convergence in the statement of Theorem 1, over a whole class of structural equation models. This can be ensured by strengthening the assumptions to hold uniformly:

(U1) The quantities in (A2)(i) and (ii) are upper-bounded by positive constants $C_1 < \infty$, $C_2 < \infty$ and $C_3 < \infty$.

(U2) The error variances in (A3) are lower bounded by a finite constant $L > 0$.

(U3) The approximation in (A4) holds uniformly over a class of functions $\mathcal{F}$: for any compact set $\mathcal{C}$ and any j, k:

$$\sup_{f^0 \in \mathcal{F}} \mathbb{E}[(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k))^2 I(X_k \in \mathcal{C})] = o(1)$$

(U4) The constant $\xi_p \geq B > 0$ for some finite constant $B > 0$.

Denote the class of distributions in an additive SEM which satisfy (U1)-(U4) by $\mathcal{P}(C_1, C_2, C_3, L, \mathcal{F}, B)$. We then obtain a uniform convergence result

$$\inf_{P \in \mathcal{P}(C_1, C_2, C_3, L, \mathcal{F}, B)} \mathbb{P}_P[\hat{\pi} \in \Pi^0] \to 1 \qquad (n \to \infty). \tag{2.14}$$

This can be shown exactly along the lines of the proof of Theorem 1 in the supplement to Bühlmann et al. (2014).

Misspecified error distribution and biased function estimation

Theorem 1 generalizes to the situation where the model in (2.3) is misspecified and the truth has independent, non-Gaussian errors $\varepsilon_1, \ldots, \varepsilon_p$ with $\mathbb{E}[\varepsilon_j] = 0$. As in Theorem 1, we make the assumption $\xi_p > 0$ in (2.9):

its justification, however, is somewhat less backed up because the identifiability results from Peters et al. (2014) and Lemma 3 do not carry over immediately. The latter results say that the set of correct orderings Π^0 can be identified from the distribution of $X_1, \ldots, X_p$, but we require in (2.9) that identifiability is given in terms of all the error variances, that is, involving only second moments. It is an open problem whether (or for which subclass of models) identifiability from the distribution carries over to automatically ensure that $\xi_p > 0$ in (2.9).

Furthermore, assume that the number of basis functions a_n for functions in $\mathcal{F}_n$ is small such that assumption (A4) does not hold, for example, $a_n = O(1)$. We denote by

$$(\sigma_j^{\pi,0,a_n})^2 = \min_{g_j \in \mathcal{F}_n^{\oplus j-1}} \mathbb{E}_{\theta^0} \left[\left(X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi) \right)^2 \right],$$

which is larger than $(\sigma_j^{\pi,0})^2$ in (2.7). Instead of (2.9), we then consider

$$\xi_p^{a_n} := \min_{\pi \notin \Pi^0,\ \pi^0 \in \Pi^0} p^{-1} \sum_{j=1}^{p} (\log(\sigma_j^{\pi,0,a_n}) - \log(\sigma_j^{\pi^0,0,a_n})). \tag{2.15}$$

Requiring

$$\liminf_{n \to \infty} \xi_p^{a_n} > 0$$

is still reasonable: if (2.9) with $\xi_p > 0$ holds because of nonlinearity of the additive functions (Peters et al., 2014), and see the interpretation above for non-Gaussian errors, we believe that it typically also holds for the best projected additive functions in $\mathcal{F}_n^{\oplus}$ as long as some nonlinearity is present when using a_n basis functions; here, the best projected additive function for the j-th variable X_j^π is defined as $f_{n;j}^\pi = \operatorname{argmin}_{g_j \in \mathcal{F}_n^{\oplus j-1}} \mathbb{E}[(X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi))^2]$. We also note that for $a_n \to \infty$, even when diverging very slowly, and assuming (A4) we have that $\xi_p^{a_n} \to \xi_p$ and thus $\liminf_{n\to\infty} \xi_p^{a_n} > 0$. In general, the choice of the number of basis functions a_n is a trade-off between identifiability (due to nonlinearity) and estimation accuracy: for a_n small we might have a smaller value in (2.15), that is, it might be that $\xi_p^{a_n} \leq \xi_p^{a_n'}$ for $a_n \leq a_n'$, which makes identifiability harder but exhibits less variability in estimation; and vice versa. In particular, the trade-off between identifiability and variance might be rather different than the classical bias-variance trade-off with respect to prediction in classical function estimation. A low complexity (with a_n small) might be better than a prediction optimal number of basis functions.

Theorem 2 below establishes the consistency for order estimation in an additive structural equation model with potentially non-Gaussian errors, even when the expansion for function estimation is truncated at few basis functions.

Theorem 2. *Consider an additive structural equation model as in* (2.3) *but with independent potentially non-Gaussian errors* $\varepsilon_1, \ldots, \varepsilon_p$ *having* $\mathbb{E}[\varepsilon_j] = 0$ $(j = 1, \ldots, p)$. *Assume either of the following:*

1. *(A1)-(A4) hold, and* $\xi_p > 0$ *in formula* (2.9) *(see also Remark 2).*
2. *(A1)-(A3) hold, and* $\liminf_{n\to\infty} \xi_p^{a_n} > 0$ *in formula* (2.15).

Then,

$$\mathbb{P}[\hat{\pi} \in \Pi^0] \to 1 \qquad (n \to \infty).$$

A proof is given in the supplement to Bühlmann et al. (2014). Again, as appearing in the discussion of Theorem 1, one can obtain uniform convergence by strengthening the assumptions to hold uniformly over a class of distributions.

2.4.2 Restricted MLE for sparse high-dimensional settings

We consider here the restricted MLE in (2.13) and show that it can cope with high-dimensional settings where $p \gg n$.

The model in (2.1) is now assumed to change with sample size n: the dimension is $p = p_n$ and the parameter $\theta = \theta_n$ depends on n. We consider the limit as $n \to \infty$ allowing diverging dimension $p_n \to \infty$ where $p_n \gg n$. For notational simplicity, we often drop the sub-index n.

We make a few additional assumptions. When fitting an additive model of X_j versus all other variables $X_{\{-j\}}$, the target of such an estimation is the best approximating additive function:

$$\mathbb{E}_{\text{add}}[X_j | X_{\{-j\}}] = \sum_{k \in \{-j\}} h_{jk}^*(X_k),$$

$$\{h_{jk}^*;\ k \in \{-j\}\} = \operatorname*{argmin}_{h_j \in \mathcal{F}^{\oplus p-1}} \mathbb{E}\left[\left(X_j - \sum_{k \in \{-j\}} h_{jk}(X_k)\right)^2\right].$$

In general, some variables are irrelevant, and we denote the set of relevant variables by A_j: $A_j \subseteq \{1, \ldots, p\} \setminus j$ is the (or a) smallest set[4] such that

$$\mathbb{E}_{\text{add}}[X_j | X_{\{-j\}}] = \mathbb{E}_{\text{add}}[X_j | X_{A_j}].$$

We assume the following:

(B1) For all $j = 1, \ldots, p$: for all $k \in \text{pa}(j)$,

$$\mathbb{E}_{\text{add}}[(X_j - \mathbb{E}_{\text{add}}[X_j | X_{A_j \setminus k}]) | X_k] \not\equiv 0.$$

Assumption (B1) requires that for each $j = 1, \ldots, p$: X_k ($k \in \text{pa}(j)$) has an additive influence on X_j given all additive effects from $X_{A_j \setminus k}$.

Lemma 4. *Assume that (B1) holds. Then, for all $j = 1, ..., p$,*

$$\text{pa}(j) \subseteq A_j.$$

A proof is given in the supplement to Bühlmann et al. (2014). Lemma 4 justifies, for the population case, to pursue preliminary neighborhood selection followed by restricted maximum likelihood estimation: as $\text{pa}(j) \subseteq A_j$, the restriction in the maximum likelihood estimator is appropriate and a true permutation in $\pi^0 \in \Pi^0$ leads to a valid restriction $R_{\pi^0, j} \supseteq \text{pa}(\pi^0(j))$ (when defined with the population sets A_j).

For estimation, we assume the following:

(B2) The selected variables in $\hat{A}_j$ from neighborhood selection satisfy: with probability tending to 1 as $n \to \infty$,

(i) $\hat{A}_j \supseteq A_j$ $(j = 1, \ldots, p)$,

(ii) $\max_{j=1,\ldots,p} |\hat{A}_j| \leq M < \infty$ for some positive constant $M < \infty$.

Assumption (B2)(i) is a rather standard screening assumption. It holds for the Group Lasso with sparsity-smoothness penalty: using a basis expansion as in (2.4), the condition is implied by a sparsity assumption, a group compatibility condition (for the basis vectors), and a beta-min condition about the minimal size of the ℓ_2-norm of the coefficients for the basis functions of the active variables in A_j; see Theorem 8.2 in Chapter 5.6 of Bühlmann and van de Geer (2011). The sparsity and group

[4] Uniqueness of such a set is not a requirement but implicitly ensured by the compatibility condition and sparsity which we invoke to guarantee B2(ii).

compatibility condition ensure identifiability of the active set and hence, they exclude concurvity (or collinearity) among the additive functions in the structural equation model. Assumption (B2)(ii) can be ensured by assuming $\max_j |A_j| \leq M_1 < \infty$ for some positive constant $M_1 < \infty$ and, for example, group restricted eigenvalue assumptions for the design matrix (with the given basis); see van de Geer et al. (2011) and Zhang and Huang (2008) for the case without groups.

Finally, we need to strengthen assumptions (A2) and (A3).

(B3) (i) For $B \subseteq \{1, \ldots, p\} \setminus j$ with $|B| \leq M$, with M as in (B2), denote by $h^B_{j,g} = (X_j - \sum_{k \in B} g_k(X_k))^2$. For some $0 < K < \infty$, it holds that

$$\max_{j=1,\ldots,p} \max_{\substack{B \subseteq \{1,\ldots,p\} \setminus j \\ |B| \leq M}} \sup_{g \in \mathcal{F}^{\oplus |B|}} \rho_K(h^B_{j,g}) \leq D_1 < \infty,$$

where

$$\rho^2_K(h^B_{j,g}) = 2K^2 \mathbb{E}_{\theta^0}[\exp(|h^B_{j,g}(X)|/K) - 1 - |h^B_{j,g}(X)|/K].$$

(ii) For $V = 1/\alpha$,

$$\max_{j=1,\ldots,p} \left(\sum_{m=1}^{\infty} (M_m^2 \mathbb{P}[X_j \in I_m])^{V/(V+4)} \right)^{(V+4)/8} \leq D_2 < \infty.$$

This assumption is typically weaker than what we require in (B3)(i), when assuming that the values M_m are reasonable (e.g., bounded).

(iii)

$$\max_j \mathbb{E}|X_j|^4 \leq D_3 < \infty,$$
$$\max_j \sup_{f \in \mathcal{F}} \mathbb{E}|f(X_j)|^4 \leq D_4 < \infty.$$

(B4) The error variances satisfy $\min_\pi \min_j (\sigma_j^{\pi,0})^2 \geq L > 0$.

Assumption (B3)(i) requires exponential moments. We note that the sum of additive functions over the set B is finite. Thus, we essentially require exponential moments for the square of finite sums of additive functions.

Theorem 3. *Consider an additive structural equation model as in* (2.3) *with independent potentially non-Gaussian errors* $\varepsilon_1, \ldots, \varepsilon_p$ *with* $\mathbb{E}[\varepsilon_j] = 0$ $(j = 1, \ldots, p)$. *Assume either of the following:*

1. *(A1), (A4) and (B1)-(B4) hold, and for* ξ_p *in* (2.9) *(see also Remark 2):*

$$\max\left(\sqrt{\log(p)/n}, \max_{j,k} \mathbb{E}[(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k))^2]\right) = o(\xi_p).$$

2. *(A1) and (B1)-(B4) hold, and for* $\xi_p^{a_n}$ *in* (2.15)*:*

$$\max\left(\sqrt{\log(p)/n}, \max_{j,k} \mathbb{E}[(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k))^2]\right) = o(\xi_p^{a_n}).$$

Then, for the restricted maximum likelihood estimator in (2.13)*:*

$$\mathbb{P}[\hat{\pi} \in \Pi^0] \to 1 \qquad (n \to \infty).$$

A proof is given in the supplement to Bühlmann et al. (2014). The assumption that $\mathbb{E}[(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k))^2]$ is of sufficiently small order can be ensured by the following condition.

($\mathbf{B_{add}}$) Consider the basis functions $b_r(\cdot)$ appearing in $\mathcal{F}_n$: for the true functions $f^0_{j,k} \in \mathcal{F}$, we assume an expansion

$$f^0_{j,k}(x) = \sum_{r=1}^{\infty} \alpha_{f^0_{j,k};r} b_r(x)$$

with smoothness condition

$$\sum_{r=k}^{\infty} |\alpha^0_{f^0_{j,k};r}| \le Ck^{-\beta}.$$

Assuming ($\mathrm{B_{add}}$) we have that $\mathbb{E}[(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k))^2] = O(a_n^{-(\beta-1-\kappa)})$ for any $\kappa > 0$: for example, when using $a_n \to \infty$, it holds for $\beta > 1$ that $\mathbb{E}[(f^0_{j,k}(X_k) - f^0_{n;j,k}(X_k))^2] \to 0$.

Uniform convergence can be obtained exactly as described after the discussion of Theorem 1: when requiring the additional uniform versions (U3)-(U4) (since (B3) and (B4) invoke already uniform bounds we do not need (U1) and (U2)), and requiring uniform convergence of the probability in (B2), we obtain uniform convergence over the corresponding class of distributions analogously as in (2.14).

2.5 Computation and implementation

In Section 2.2 we have decomposed the problem of learning DAGs from observational data into two main parts: finding the correct order (Section 2.2.4) and feature selection (Section 2.2.5). Our algorithm and implementation consists of two corresponding parts: *IncEdge* is a greedy procedure providing an estimate $\hat{\pi}$ for equation (2.11) and *Prune* performs the feature selection. Section 2.3.1 discusses the benefits of performing a preliminary neighborhood selection before estimating the causal order, and we call the corresponding part *PNS*. The combination *PNS* + *IncEdge* provides an estimate for equation (2.13).

The three components of our implementation are described in the following sections, Figures 2.1, 2.2 and 2.3 present the steps graphically. We regard the modular structure of the implementation as an advantage; each of the three parts could be replaced by an alternative method (as indicated in the subsections below).

The code for CAM is provided in the `R`-package `CAM`.

2.5.1 Preliminary Neighborhood Selection: *PNS*

As described in Section 2.3.1 we fit an additive model for each variable X_j against all other variables $X_{\{-j\}}$. We implement this with a boosting method for additive model fitting (Bühlmann and Hothorn, 2007; Bühlmann and Yu, 2003), using the R-function `gamboost` from the package `mboost` (Hothorn et al., 2010). We select the ten variables that have been picked most often during 100 iterations of the boosting method; hereby, we only consider variables that have been picked at least three times during the iterations. The sets $\hat{A}_j$ obtained by this procedure estimate $A_j \supseteq \mathrm{pa}(j)$ as shown in Lemma 4. We construct a graph in which for each j, the set $\hat{A}_j$ is the parental set for node j corresponding to the variable X_j. Figure 2.1 summarizes this step. We say that the set of "possible parents" of node j has been reduced to the set $\hat{A}_j$. Importantly, we do not disregard true parents if the sample size is large enough (Section 2.4.2, Lemma 4).

Instead of the boosting method, we could alternatively use additive model fitting with a sparsity- or sparsity-smoothness penalty (Meier et al., 2009; Ravikumar et al., 2009).

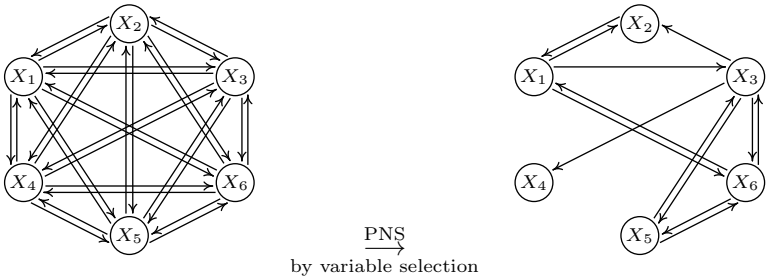


Figure 2.1: Step *PNS*. For each variable the set of possible parents is reduced (in this plot, a directed edge from X_k to X_j indicates that X_k is a selected variable in $\hat{A}_j$ and a possible parent of X_j). This reduction leads to a considerable computational gain in the remaining steps of the procedure.

2.5.2 Estimating the correct order by greedy search: *IncEdge*

Let us first consider the situation without *PNS*. Searching over all permutations π for finding $\hat{\pi}$ in (2.11) is computationally infeasible if the number of variables p is large. We propose a greedy estimation procedure that starts with an empty DAG and adds at each iteration the edge $k \to j$ between nodes k and j that corresponds to the largest gain in log-likelihood. We therefore compute the score function in (2.11), with D corresponding to the current DAG,

$$\sum_{j=1}^{p} \log(\hat{\sigma}_j^D) = \sum_{j=1}^{p} \log\left(\left\| X_j - \sum_{k \in \mathrm{pa}_D(j)} \hat{f}_{j,k}^D(X_k) \right\|_{(n)} \right)$$

and construct a matrix, whose entry (k, j) specifies by how much this score is reduced after adding the edge $k \to j$ and, therefore, allowing a non-constant function $f_{j,k}$ (see Figure 2.2). For implementation, we employ additive model fitting with penalized regression splines (with ten basis functions per variable), using the R-function `gam` from the `R`-package `mgcv`, in order to obtain estimates $\hat{f}_{j,k}$ and $\hat{\sigma}_j$. After the addition of an edge, we only need to recompute the jth column of the score matrix (see Figure 2.2) since the score decomposes over all nodes. In order to avoid cycles we remove further entries of the score matrix. After $p(p-1)/2$ iterations the graph has been completed to a fully connected DAG. The latter cor-

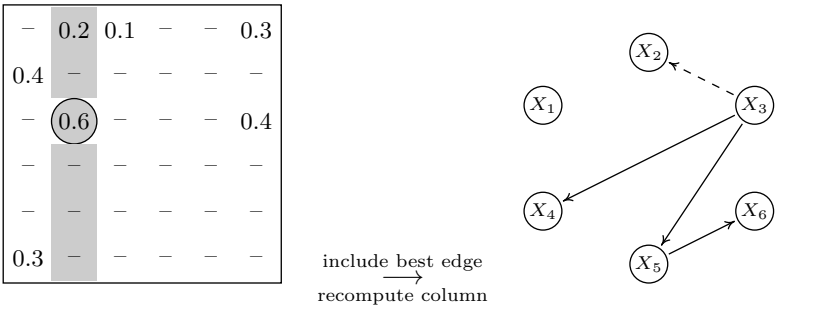


Figure 2.2: Step *IncEdge*. At each iteration the edge leading to the largest decrease of the negative log-likelihood is included.

responds to a unique permutation $\hat{\pi}$. This algorithm is computationally rather efficient and can easily handle graphs of up to 30 nodes without *PNS* (see Section 2.6.2).

If we have performed *PNS* as in Section 2.5.1 we sparsify the score matrix from the beginning. We only consider entries (k, j) for which k is considered to be a possible parent of j. This way the algorithm is feasible for up to a few thousands of nodes (see Section 2.6.3).

Alternative methods for (low-dimensional) additive model fitting include backfitting (cf. Mammen and Park, 2006).

2.5.3 Pruning of the DAG by feature selection: *Prune*

Section 2.2.5 describes sparse regression techniques for pruning the DAG that has been estimated by Step *IncEdge*, see Figure 2.3. We implement this task by applying significance testing of covariates, based on the `R`-function `gam` from the `R`-package `mgcv` and declaring significance if the reported p-values are lower or equal to 0.001, independently of the sample size (for problems with small sample size, the p-value threshold should be increased).

If the DAG estimated by (*PNS* and) *IncEdge* is a super-DAG of the true DAG, the estimated interventional distributions are correct, see Section 2.2.6. This does not change if *Prune* removes additional "superfluous" edges. The structural Hamming distance to the true graph, however, may reduce significantly after performing *Prune*, see Section 2.6.2. Alternative

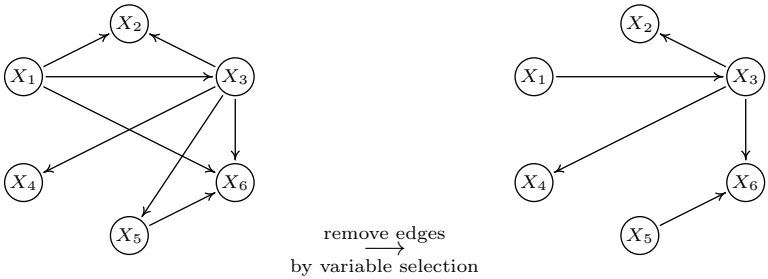


Figure 2.3: Step *Prune*. For each node, variable selection techniques are exploited to remove non-relevant edges.

methods for hypothesis testing in (low-dimensional) additive models are possible (cf. Wood, 2006), or one could use penalized additive model fitting for variable selection (Meier et al., 2009; Ravikumar et al., 2009; Yuan and Lin, 2006).

2.6 Numerical results for simulated data

We show the effectiveness of each step in our algorithm (Section 2.6.2) and compare the whole procedure to other state-of-the-art methods (Section 2.6.3). We investigate empirically the role of non-injective functions (Section 2.6.4) and discuss the linear Gaussian case (Section 2.6.5). In Section 2.6.6 we further check the robustness of our method against model misspecification, that is, in the case of non-Gaussian noise or non-additive functions. For evaluation we compute the structural intervention distance that we introduce in Section 2.6.1.

For simulating data, we randomly choose a correct ordering π^0 and connect each pair of variables (nodes) with a probability p_c. If not stated otherwise, each of the possible $p(p-1)/2$ connections is included with a probability of $p_c = 2/(p-1)$ resulting in a sparse DAG with an expected number of p edges. Given the structure, we draw the functions $f_{j,k}$ from a Gaussian process with a Gaussian (or RBF) kernel with bandwidth one and add Gaussian noise with standard deviation uniformly sampled between $1/5$ and $\sqrt{2}/5$. All nodes without parents have a standard deviation between 1 and $\sqrt{2}$. The experiments are based on 100 repetitions if the description does not say differently.

2.6.1 Structural Intervention Distance

As a performance measure, we consider the recently proposed structural intervention distance (SID), see Peters and Bühlmann (2015). The SID is well-suited for quantifying the correctness of an order among variables, mainly in terms of inferring causal effects afterwards. It counts the number of wrongly estimated causal effects. Thus, the SID between the true DAG D^0 and the fully connected DAGs corresponding to the true permutations $\pi^0 \in \Pi^0$ is zero, see Section 2.2.6.

2.6.2 Effectiveness of preliminary neighborhood selection and pruning

We demonstrate the effect of the individual steps of our algorithm. Figure 2.4 shows the performance (in terms of SID and SHD) of our method and the corresponding time consumption (using eight cores) depending on which of the steps are performed.

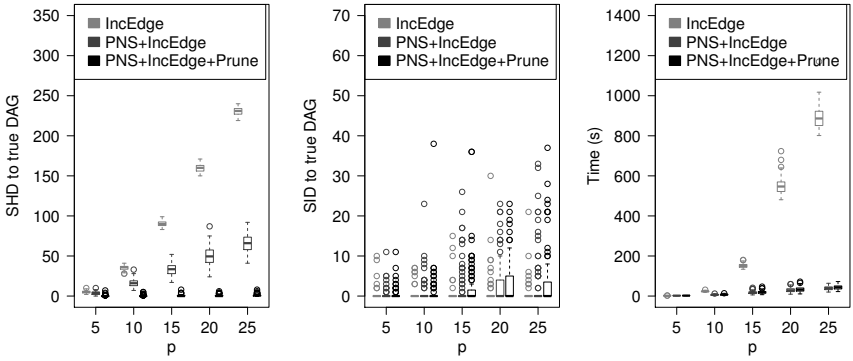


Figure 2.4: The plots show the effect of the individual steps of our method. *Prune* reduces the SHD to the true DAG but leaves the SID almost unchanged. *PNS* reduces the computation time, especially for large p.

If only *IncEdge* is used, the SHD is usually large because the output is a fully connected graph. Only after the Step *Prune* the SHD becomes small. As discussed in Section 2.2.6 the pruning does not make a big difference for the SID. Performing these two steps is not feasible for large p. The

time consumption is reduced significantly if we first apply the preliminary neighborhood selection *PNS*. In particular, this first step is required in the case of $p > n$ in order to avoid a degeneration of the score function.

2.6.3 Comparison to existing methods

Different procedures have been proposed to address the problem of inferring causal graphs from a joint observational distribution. We compare the performance of our method to greedy equivalence search (GES) (Chickering, 2002), the PC algorithm (Spirtes et al., 2000), the conservative PC algorithm (CPC) (Ramsey et al., 2006), LiNGAM (Shimizu et al., 2006) and regression with subsequent independence tests (RESIT) (Mooij et al., 2009; Peters et al., 2014). The latter has been used with a significance level of $\alpha = 0$, such that the method does not remain undecided. Both PC methods are equipped with $\alpha = 0.01$ and partial correlation as independence test. GES is used with a linear Gaussian score function. Thus, only RESIT is able to model the class of nonlinear additive functions. We apply the methods to DAGs of size $p = 10$ and $p = 100$, whereas in both cases, the sample size is $n = 200$. RESIT is not applicable for graphs with $p = 100$ due to computational reasons. Figure 2.5 shows that our proposed method outperforms the other approaches both in terms of SID and SHD.

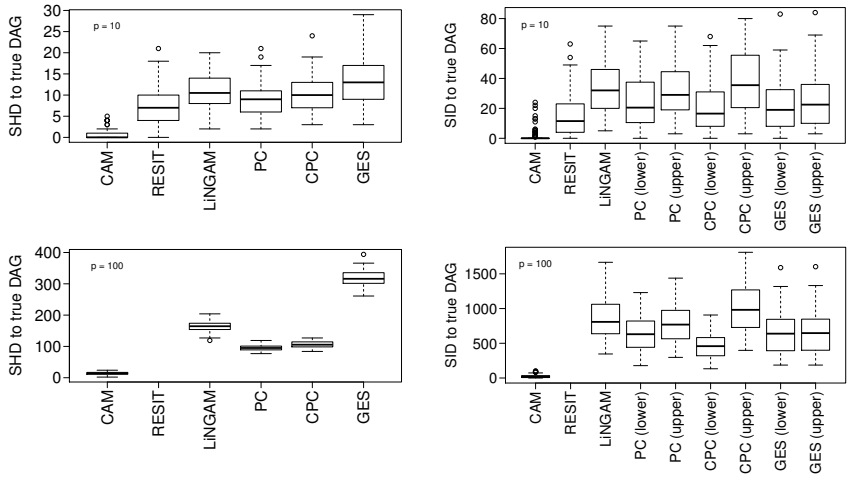


Figure 2.5: SHD (left) and SID (right) for different methods on sparse DAGs with $p = 10$ (top) and $p = 100$ (bottom); the sample size is $n = 200$.

The difference between the methods becomes even larger for dense graphs with an expected number of $4p$ edges and strong varying degree of nodes (results not shown).

Only the PC methods and the proposed method CAM scale to high-dimensional data with $p = 1000$ and $n = 200$. Keeping the same (sparse) setting as above results in SHDs of 1214 ± 37, 1330 ± 40 and 477 ± 19 for PC, CPC and CAM, respectively. These results are based on five experiments.

2.6.4 Injectivity of model functions

In general, the nonlinear functions that are generated by Gaussian processes are not injective. We therefore test CAM for the case where every function in (2.1) is injective. Correct direction of edges (j, k) is a more difficult task in this setting. We sample sigmoid-type functions of the form

$$f_{j,k}(x_k) = a \cdot \frac{b \cdot (x_k + c)}{1 + |b \cdot (x_k + c)|}$$

with $a \sim \text{Exp}(4) + 1$, $b \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathcal{U}([-2, 2])$; as before, we use Gaussian noise. Note that some of these functions may be very close to linear functions which makes the direction of the corresponding edges difficult to identify. Figure 2.6 shows a comparison of the performance of CAM in the previously applied setting with Gaussian processes and in the new setting with sigmoid-type functions. As expected, the performance of CAM decreases in this more difficult setting but is still better than for the competitors such as RESIT, LiNGAM, PC, CPC and GES (not shown).

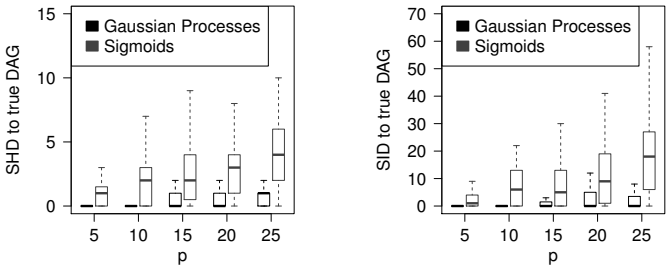


Figure 2.6: SHD (left) and SID (right) for various values of p and $n = 300$. The plots compare the performances of CAM for the additive SEM (2.1) with functions generated by Gaussian processes (non-injective in general) and sigmoid-type functions (injective).

2.6.5 Linear Gaussian SEMs

In the linear Gaussian setting, we can only identify the Markov equivalence class of the true graph (if we assume faithfulness). We now sample data from a linear Gaussian SEM and expand the DAGs that are estimated by CAM and LiNGAM to CPDAGs, that is, we consider the corresponding Markov equivalence classes. The two plots in Figure 2.7 compare the different methods for $p = 10$ variables and $n = 200$. They show the structural Hamming distance (SHD) between the estimated and the true Markov equivalence class (left), as well as lower and upper bounds for the SID (right). By the definition of lower and upper bounds of the SID, the SID between the true and estimated DAG lies in between those values. The proposed method has a disadvantage because it uses nonlinear regression instead of linear regression. The performance is nevertheless comparable. Remark 1 discusses that at least in principle, this scenario is detectable.

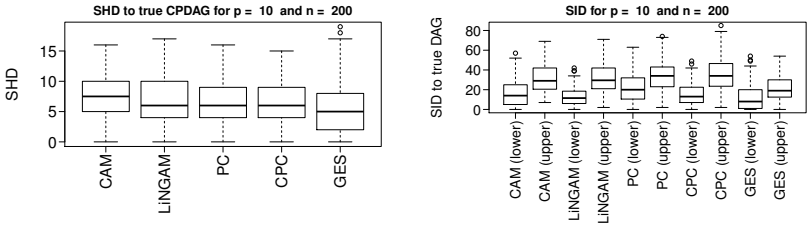


Figure 2.7: Comparison to existing methods for data generated by linear Gaussian SEM. SHD between true and estimated CPDAG (left), lower and upper bounds for SID between true DAG and estimated CPDAG (right).

2.6.6 Robustness against non-additive functions and non-Gaussian errors

All of Chapter 2 focuses on the additive model (2.1) and Gaussian noise. The score functions (2.11) and (2.13) and their corresponding optimization problems depend on these model assumptions. The DAG remains identifiable (under weak assumptions) even if the functions of the data generating process are not additive or the noise variables are non-Gaussian (cf. Peters et al., 2014). The following experiments analyze the empirical performance

of our method under these misspecifications. The case of misspecified error distributions is discussed in Section 2.4.1.

As a first experiment we examine deviations from the Gaussian noise assumption by setting $\varepsilon_j = \text{sign}(N_j)|N_j|^\gamma$ with $N_j \sim \mathcal{N}(0, \sigma_j^2)$ for different exponents $0.1 \leq \gamma \leq 4$. Only $\gamma = 1$ corresponds to normally distributed noise. Figure 2.8 shows the change in SHD and SID when varying γ.

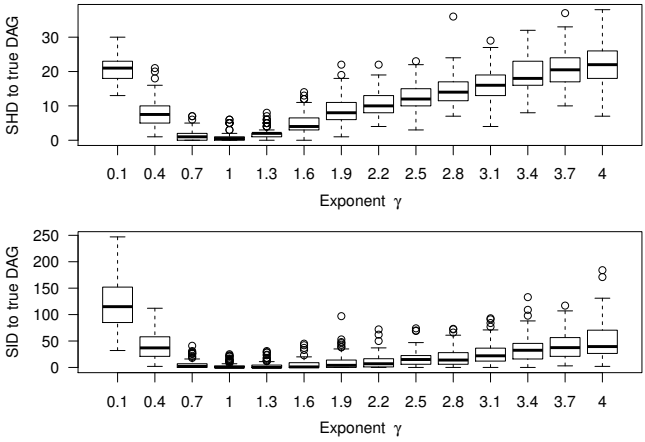


Figure 2.8: SHD (top) and SID (bottom) for $p = 25$ and $n = 300$ in the case of misspecified models. The plot shows deviations of the noise from a normal distribution (only $\gamma = 1$ corresponds to Gaussian noise).

As a second experiment, we examine deviations from additivity by simulating from the model

$$X_j = \omega \cdot \sum_{k \in \text{pa}_D(j)} f_{j,k}(X_k) + (1 - \omega) \cdot f_j(X_{\text{pa}_D(j)}) + \varepsilon_j$$

for different values of $\omega \in [0, 1]$ and Gaussian noise. Both, $f_{j,k}$ and f_j are drawn from a Gaussian process using an RBF kernel with bandwidth one. Note that $\omega = 1$ corresponds to the fully additive model (2.3), whereas for $\omega = 0$, the value of X_j is given as a non-additive function of all its parents. Figure 2.9 shows the result for a sparse truth with expected number of p edges (top) and a non-sparse truth with expected number of $4p$ edges (lower). In sparse DAGs, many nodes have a small number of

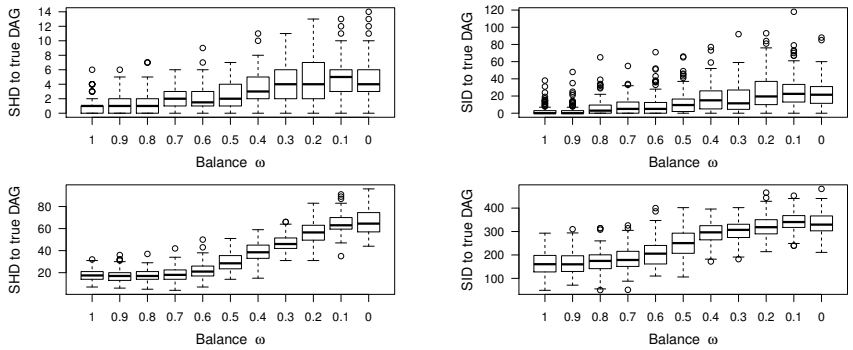


Figure 2.9: SHD (left) and SID (right) for $p = 25$ and $n = 300$ in the case of misspecified models. The plot shows deviations from additivity for sparse (top) and non-sparse (bottom) truths, respectively (only $\omega = 1$ corresponds to a fully additive model).

parents and our algorithm yields a comparably small SID even if the model contains non-additive functions. If the underlying truth is non-sparse, the performance of our algorithm becomes worse but it is still slightly better than PC which achieves average lower bounds of SID values of roughly 520, both for $\omega = 1$ and for $\omega = 0$ (not shown).

2.7 Real data application

We apply our methodology to microarray data described in Wille et al. (2004). The authors concentrate on 39 genes (118 observed samples) on two isoprenoid pathways in Arabidopsis thaliana. The dashed edges in Figures 2.10 and 2.11 indicate the causal direction within each pathway. While graphical Gaussian models are applied to estimate the underlying interaction network by an undirected model in Wille et al. (2004), our CAM procedure estimates the structure by a directed acyclic graph.

Given a graph structure, we can compute p-value scores as described in Section 2.5.3. Figure 2.10 shows the twenty best scoring edges of the graph estimated by our proposed method CAM (the scores should not be interpreted as p-values anymore since the graph has been estimated from data). We also apply stability selection (Meinshausen and Bühlmann, 2010) to this data set. We therefore consider 100 different subsamples of size 59

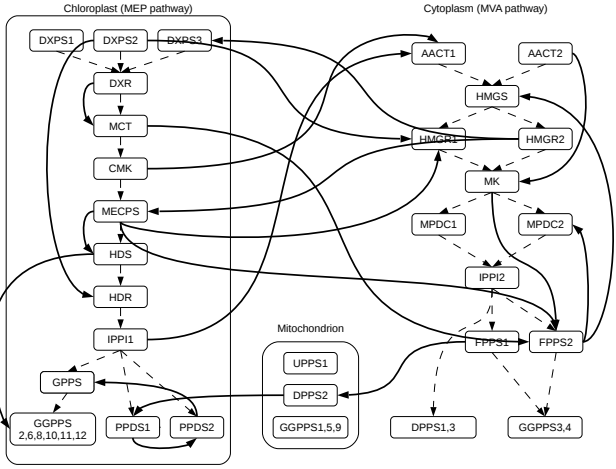


Figure 2.10: Gene expressions in isoprenoid pathways. The twenty best scoring edges provided by the method CAM.

and record the edges that have been considered at least 57 times as being among the 20 best scoring edges. Under suitable assumptions this leads to an expected number of false positives being less than two (Meinshausen and Bühlmann, 2010). These edges are shown in Figure 2.11. They connect genes within one of the two pathways and their directions agree with the overall direction of the pathways. Our findings are therefore consistent with the prior knowledge available. The link MCT $\rightarrow$ CMK does not appear in Figure 2.10 since it was ranked as the 22nd best scoring edge.

2.8 Conclusions and extensions

2.8.1 Conclusions

We have proposed maximum likelihood estimation and its restricted version for the class of additive structural equation models (i.e., causal additive models, CAMs) with Gaussian errors where the causal structure (underlying DAG) is identifiable from the observational probability distribution (Peters et al., 2014). A key component of our approach is to

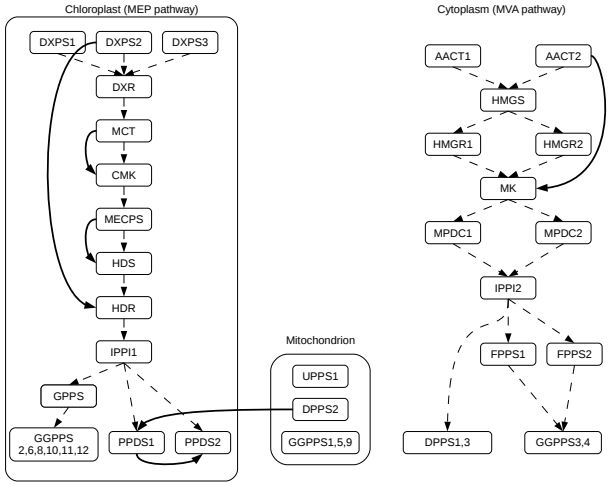


Figure 2.11: Gene expressions in isoprenoid pathways. Edges estimated by stability selection: all directions are in correspondence with the direction of the pathways.

decouple order search among the variables from feature or edge selection in DAGs. Regularization is only necessary for the latter while estimation of an order can be done with a non-regularized (restricted) maximum likelihood principle. Thus, we have substantially simplified the problem of structure search and estimation for an important class of causal models. We established consistency of the (restricted) maximum likelihood estimator for low- and high-dimensional scenarios, and we also allow for misspecification of the error distribution. Furthermore, we developed an efficient computational algorithm which can deal with many variables, and the new method's accuracy and performance is illustrated with a variety of empirical results for simulated and real data. We found that we can do much more accurate estimation for identifiable, nonlinear CAMs than for non-identifiable linear Gaussian structural equation models.

2.8.2 Extensions

The estimation principle of first pursuing order search based on non-regularized maximum likelihood and then using penalized regression for

feature selection works with other structural equation models where the underlying DAG is identifiable from the observational distribution. Closely related examples include nonlinearly transformed additive structural equation models (Zhang and Hyvärinen, 2009) or Gaussian structural equation models with same error variances (Peters and Bühlmann, 2014).

If the DAG D is non-identifiable from the distribution P, the methodology needs to be adapted; see also Remark 1 considering the linear Gaussian SEM. The true orders Π^0 can be defined as the set of permutations which lead to most sparse autoregressive representations as in (2.5): assuming faithfulness, these orders correspond to the Markov equivalence class of the underlying DAG. Therefore, for estimation, we should use regularized maximum likelihood estimation leading to sparse solutions with, for example, the ℓ_0-penalty (Chickering, 2002; van de Geer and Bühlmann, 2013).

Finally, it would be very interesting to extend (sparse) permutation search to (possibly non-identifiable) models with hidden variables (Colombo et al., 2012; Janzing et al., 2009; Pearl, 2000; Spirtes et al., 2000) or with graph structures allowing for cycles (Mooij et al., 2011; Mooij and Heskes, 2013; Richardson, 1996; Spirtes, 1995). Note that unlike linear Gaussian models, CAMs are not closed under marginalization: if X, Y and Z follow a CAM (2.1), then X and Y do not necessarily remain in the class of CAMs.

Chapter 3

Identifiability and estimation of partially linear additive models[1]

We consider the identifiability and estimation of partially linear additive structural equation models with Gaussian noise (PLSEMs). Existing identifiability results in the framework of additive SEMs with Gaussian noise are limited to linear and nonlinear SEMs, which can be considered as special cases of PLSEMs with vanishing nonparametric or parametric part, respectively. We close the wide gap between these two special cases by providing a comprehensive theory of the identifiability of PLSEMs by means of (A) a graphical, (B) a transformational, (C) a functional and (D) a causal ordering characterization of PLSEMs that generate a given distribution $\mathbb{P}$. In particular, the characterizations (C) and (D) answer the fundamental question to which extent nonlinear functions in additive SEMs with Gaus-

[1]This chapter is a slightly modified version of the preprint Ernest, J., Rothenhäusler, D., and Bühlmann, P. (2016). *Causal inference in partially linear structural equation models: identifiability and estimation.* arXiv:1607.05980. Jan Ernest and Dominik Rothenhäusler are shared first authors and contributed equally to this work. The main contributions of Jan Ernest are the graphical and transformational characterizations (the proof of the transformational characterization was mostly done by D.R.), the conceptual development and implementation of the estimation procedures, major parts of the proofs of their correctness and consistency, and the realization of all simulation experiments. Jan Ernest wrote the main text (excluding Sections 3.2.2 and 3.2.3).

sian noise restrict the set of potential causal models and hence influence the identifiability. On the basis of the transformational characterization (B) we provide a score-based estimation procedure that outputs the graphical representation (A) of the distribution equivalence class of a given PLSEM. We derive its (high-dimensional) consistency and demonstrate its performance on simulated datasets.

3.1 Introduction

Causal inference is relevant in many scientific disciplines. Examples are the identification of causal molecular mechanisms in genomics (Statnikov et al., 2012; Stekhoven et al., 2012), the investigation of causal relations among activity in brain regions from fMRI data (Ramsey et al., 2010) or the search for causal associations in public health (Glass et al., 2013).

A major research topic in causal inference aims at establishing causal dependencies based on purely observational data. The notion "observational" commonly refers to the fact that one obtains the data from the system of variables under consideration without subjecting it to external manipulations. Typically, one then assumes that the observed data has been generated by an underlying causal model and tries to draw conclusions about its structure.

Two main research tasks in this setting are the identifiability and estimation of the underlying causal model. In this chapter we address both of them for partially linear additive structural equation models with Gaussian noise (PLSEMs). Unlike in regression where partially linear models are mainly studied because of efficiency gains in estimation, the use of partially linear models has a deeper meaning in causal inference. In fact, as we will show, it is closely connected to identifiability. The functional form of an additive component directly influences the identifiability of the corresponding (and also other) causal relations. For this reason we strongly believe that the understanding of the identifiability of PLSEMs is important. First and foremost, it raises the awareness of potentially limited (or increased) identifiability in the presence of linear (or nonlinear) relations in the data. Second, by not restricting the functions to be either all linear or all nonlinear, PLSEMs allow for a flexible modeling approach.

We start by reviewing and introducing important concepts in Section 3.1.1. We then provide a brief overview of related work in Section 3.1.2 and

explicitly state the main contributions of this work in Section 3.1.3.

3.1.1 Problem description and important concepts

We consider p random variables $X = (X_1, ..., X_p)$ with joint distribution $\mathbb{P}$, which is assumed to be Markov with respect to an underlying directed acyclic graph (DAG). A DAG $D = (V, E)$ is an ordered pair consisting of a set of vertices $V = \{1, ..., p\}$ associated with the variables $\{X_1, ..., X_p\}$, and a set of directed edges $E \subset V^2$ such that there are no directed cycles. A directed edge between the nodes i and j in D is denoted by $i \to j$. Node i is called a *parent* of node j and j is called a *child* of i. Moreover, the edge is said to be oriented *out of* i and *into* j. If $i \to j$ or $i \leftarrow j$, i and j are called *adjacent* and the edge is *incident* to i and j. The *degree* of a node i, denoted by $\deg_D(i)$, counts the number of edges incident to node i in DAG D. A node k that can be reached from i by following directed edges is called *descendant* of i. We use the convention that any node is a descendant of itself. The set $\mathrm{pa}_D(j) = \{i \mid i \to j \text{ in } D\}$ consists of all parents of node j. The multi-index notation $X_{\mathrm{pa}_D(j)}$ denotes the set of variables $\{X_i\}_{i \in \mathrm{pa}_D(j)}$. An edge $i \to j$ is said to be *covered* in D, if $\mathrm{pa}_D(i) = \mathrm{pa}_D(j) \setminus \{i\}$. In that case, $\mathrm{pa}_D(i)$ is a *cover* for edge $i \to j$. The process of changing the orientation of a covered edge from $i \to j$ to $i \leftarrow j$ is referred to as a *covered edge reversal*. A triple (i, j, k) is called a *v-structure*, if $\{i, j\} \subseteq \mathrm{pa}_D(k)$ and i and j are not adjacent. The graph obtained by replacing all directed edges $i \to j$ by undirected edges $i - j$ is called *skeleton* of D. The *pattern* of a DAG D is the graph with the same skeleton as D and $i \to j$ is directed if and only if it is part of a v-structure in D. A permutation $\sigma : V \to V$ is a *causal ordering* of D if $\sigma(i) < \sigma(j)$ for all $i \to j$ in D. DAGs may be used as underlying structures for structural equation models (SEMs). A SEM relates the distribution of every random variable $\{X_1, ..., X_p\}$ to the distribution of its direct causes (the parents in the corresponding DAG D) and random noise. In its most general form,

$$X_j = f_j(X_{\mathrm{pa}_D(j)}, \varepsilon_j), \qquad j = 1, ..., p, \tag{3.1}$$

where $\{f_j\}_{j=1,...,p}$ are functions from $\mathbb{R}^{|\mathrm{pa}_D(j)|+1} \to \mathbb{R}$ and $\{\varepsilon_j\}_{j=1,...,p}$ are mutually independent noise variables. Lastly, for a function $F : \mathbb{R}^p \to \mathbb{R}^p$, we write $\mathrm{D}F$ for the Jacobian of F.

Main focus: PLSEMs

In this work we study the restriction of the general SEM in equation (3.1) to *partially linear additive SEMs with Gaussian noise (PLSEMs)*:

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) + \varepsilon_j, \tag{3.2}$$

where $\mu_j \in \mathbb{R}$, $f_{j,i} \in C^2(\mathbb{R})$, $f_{j,i} \not\equiv 0$, such that $\mathbb{E}[f_{j,i}(X_i)] = 0$, and $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j^2 > 0$ for $j = 1, ..., p$. Likewise, we may write

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D^{\mathrm{L}}(j)} \alpha_{j,i} X_i + \sum_{i \in \mathrm{pa}_D^{\mathrm{NL}}(j)} f_{j,i}(X_i) + \varepsilon_j,$$

with $\alpha_{j,i} \in \mathbb{R} \setminus \{0\}$, μ_j, $f_{j,i}$, ε_j as above, $\mathrm{pa}_D^{\mathrm{L}}(j) \cup \mathrm{pa}_D^{\mathrm{NL}}(j) = \mathrm{pa}_D(j)$ and $\mathrm{pa}_D^{\mathrm{L}}(j) \cap \mathrm{pa}_D^{\mathrm{NL}}(j) = \emptyset$. Note that we do not *a priori* fix the sets $\mathrm{pa}_D^{\mathrm{L}}(j)$ and $\mathrm{pa}_D^{\mathrm{NL}}(j)$. For $\mathbb{P}$ generated by a PLSEM with DAG D, the PLSEM corresponding to D is unique (Lemma 8). Therefrom, we call an edge $i \to j$ in D a *(non-)linear edge*, if $f_{j,i}$ in the PLSEM corresponding to D is (non-)linear. Note that the concept of (non-)linearity of an edge is defined with respect to a specific DAG D. Depending on the orientations of other edges, the status of an edge $i \to j$ may change from linear to nonlinear. An example is given in Figure 3.1.

Figure 3.1: Two DAGs D_1 and D_2 with linear edges (dashed) and nonlinear edges (solid). Let us give a brief outlook: let $\mathbb{P}$ be generated by a PLSEM with DAG D_1. In this work we prove that there exists a PLSEM with DAG D_2 that generates the same distribution $\mathbb{P}$. Moreover, we show that D_1 and D_2 are the only two DAGs with a corresponding PLSEM that generates $\mathbb{P}$. For now, simply note that $1 \to 3$ is linear in D_1, but nonlinear in D_2.

The restriction to additive SEMs is interesting from both a statistical and computational perspective as the estimation of additive functions is well understood and one largely avoids the curse of dimensionality. The assumption of Gaussian noise is necessary for our theoretical results to hold. In fact, identifiability properties may deteriorate in partially linear models with arbitrary noise distributions, see Section 3.1.2. We therefore

consider PLSEMs to be among the most general SEMs with reasonable estimation properties.

Main task: characterization of all PLSEMs that generate $\mathbb{P}$

The main task of this chapter is the systematic characterization of all PLSEMs that generate a given distribution $\mathbb{P}$ under very general assumptions. In particular: how do edge functions in different PLSEMs relate to each other? How does changing a single linear edge to a nonlinear edge affect the set of potential underlying PLSEMs? Do causal orderings of different DAGs corresponding to PLSEMs that generate $\mathbb{P}$ share certain properties?

Under faithfulness, it may be natural to characterize all PLSEMs that generate $\mathbb{P}$ by their corresponding DAGs as these DAGs are restricted to a subset of the Markov equivalence class (see Section 3.1.2). For a distribution $\mathbb{P}$ that has been generated by a faithful PLSEM, we call the set of DAGs

$$\mathscr{D}(\mathbb{P}) := \left\{ D \;\middle|\; \begin{array}{l} \mathbb{P} \text{ is faithful to } D \text{ and there exists a} \\ \text{PLSEM with DAG } D \text{ that generates } \mathbb{P} \end{array} \right\}$$

the *(PLSEM) distribution equivalence class*. Can we build on characterizations of the Markov equivalence class to characterize $\mathscr{D}(\mathbb{P})$? For example, can $\mathscr{D}(\mathbb{P})$ also be graphically represented by a single PDAG? Is it possible to efficiently estimate $\mathscr{D}(\mathbb{P})$? Before we explain our approaches to answer these questions in Section 3.1.3, let us briefly summarize related work.

3.1.2 Related work

First, we discuss the identifiability of general SEMs. We then motivate why our theoretical results close a relevant "gap" by reviewing existing identifiability results for two special cases of PLSEMs where either all the functions $f_{j,i}$ are exclusively linear or exclusively nonlinear. Finally, we briefly comment on the assumption of Gaussian noise.

Identifiability of general SEMs

In the general SEM as defined in equation (3.1) one cannot draw any conclusions about D given $\mathbb{P}$ without making further assumptions. One

such assumption commonly made is faithfulness (cf. Section 3.2.1). Under faithfulness, one can identify the Markov equivalence class of D (a set of DAGs that all entail the same conditional independences), see, for example, Pearl (2009). Markov equivalence classes are well-characterized. In fact, the Markov equivalence class of a DAG D consists of all DAGs with the same skeleton and v-structures as D (Verma and Pearl, 1990) and can be graphically represented by a single partially directed graph (cf. Section 3.2.1). Moreover, any two Markov equivalent DAGs can be transformed into each other by a sequence of distinct covered edge reversals (Chickering, 1995).

The estimation of the general SEM is difficult due to the curse of dimensionality in fully nonparametric estimation. In combination with the unidentifiability, this motivates the use of restricted SEMs, which have better estimation properties and for which it is possible to achieve (partial) identifiability of the SEM (even without assuming faithfulness), see Section 3.2.2 or the paper of Peters et al. (2014) for an overview.

Special case of PLSEM: Linear Gaussian SEM

A widespread specification of PLSEMs are linear Gaussian SEMs, which have the same identifiability properties as the general SEMs: without additional assumptions they are unidentifiable, whereas under faithfulness, their distribution equivalence class equals the Markov equivalence class, see, for example, Spirtes and Zhang (2016).

The estimation of the Markov equivalence class of linear Gaussian SEMs in the low-dimensional case has been addressed in, e.g., Chickering (2002) and Spirtes et al. (2000), whereas the high-dimensional scenario (requiring sparsity of the true underlying DAG) is discussed in, e.g., Bühlmann (2013), Kalisch and Bühlmann (2007), Nandy et al. (2016), and van de Geer and Bühlmann (2013).

An exception of identifiability of linear Gaussian SEMs occurs if all ε_j have equal variances $\sigma_j^2 = \sigma^2 > 0, \forall j$. Under this assumption, the true underlying DAG D is identifiable (Peters and Bühlmann, 2014). Yet, the assumption of equal noise variances seems to be overly restrictive in many scenarios. In general, the linearity assumption may be rather restrictive if not implausible in some cases.

Special case of PLSEM: Causal additive model (CAM)

Interestingly, the assumption of exclusively nonlinear functions $f_{j,i}$ in equation (3.2) greatly improves the identifiability properties, see Hoyer et al. (2009) for the bivariate case and Peters et al. (2014) for a general treatment. In fact, if all $f_{j,i}$ are nonlinear and three times differentiable, $\mathscr{D}(\mathbb{P})$ only consists of the single true underlying DAG D (Peters et al., 2014, Corollary 31 (ii)); see also Lemma 1 in Chapter 2. The nonlinearity assumption is crucial, though. The authors provide an example where two DAGs are distributionally equivalent if one of the nonlinear functions is replaced by a linear function (Peters et al., 2014, Example 26).

Various estimation methods have been introduced for additive nonlinear SEMs to infer the underlying DAG (Nowzohour and Bühlmann, 2016; Peters et al., 2014; van de Geer, 2014). In particular, a restricted maximum likelihood estimation method called CAM (cf. Chapter 2), which is consistent in the low- and high-dimensional setting (assuming a sparse underlying DAG), has been proposed specifically for nonlinear additive SEMs with Gaussian noise (Bühlmann et al., 2014).

Importance of Gaussian noise for the identifiability of PLSEMs

The identifiability properties of linear SEMs generally improve if one allows for non-Gaussian noise distributions. In fact, if all but one ε_j are assumed to be non-Gaussian (commonly referred to as LiNGAM setting), the underlying DAG D is identifiable (Shimizu et al., 2006). A general theory for linear SEMs with arbitrary noise distributions is presented in Hoyer et al. (2008). Both papers also propose estimation procedures for the respective model classes.

Unfortunately, the situation is different for PLSEMs: identifiability can be lost if one considers PLSEMs with non-Gaussian (or arbitrary) noise distributions. This can be seen from a specific example of a bivariate linear SEM with Gumbel-distributed noise, which is identifiable in the LiNGAM framework, but for which there exists a nonlinear additive backward model (Hoyer et al., 2009). Still, this example seems to be rather particular. In fact, for bivariate additive SEMs, all unidentifiable cases of additive models can be classified into five categories, see Peters et al. (2014) and Zhang and Hyvärinen (2009).

3.1.3 Our contribution

As discussed in Section 3.1.2, there exists a wide "identifiability gap" for PLSEMs. Their identifiability has only been studied for the two special cases of linear SEMs and entirely nonlinear additive SEMs. Moreover, to the best of our knowledge, it has not yet been understood to what extent (single) nonlinear functions in additive SEMs with Gaussian noise restrict the underlying causal model. We close the "identifiability gap" for PLSEMs and answer the questions raised in Section 3.1.1 with the following theoretical results:

(A) A graphical representation of $\mathscr{D}(\mathbb{P})$ with a single partially directed graph $G_{\mathscr{D}(\mathbb{P})}$ in Section 3.2.1 (analogous to the use of CPDAGs to represent Markov equivalence classes).

(B) A transformational characterization of $\mathscr{D}(\mathbb{P})$ through sequences of covered *linear* edge reversals in Section 3.2.1 (analogous to the characterization of Markov equivalence classes via sequences of covered edge reversals in Chickering (1995)).

(C) A functional characterization of PLSEMs in Section 3.2.2: PLSEMs that generate the same distribution $\mathbb{P}$ are constant rotations of each other.

(D) A characterization of PLSEMs based on causal orderings in Section 3.2.2, which, in particular, precisely specifies to what extent nonlinear functions in PLSEMs restrict the set of potential causal orderings.

The first two characterizations hold only under faithfulness, the third and fourth are general. We will give details on the precise interplay between nonlinearity and faithfulness in Section 3.2.3. Building on the transformational characterization result in (B) we provide an efficient score-based estimation procedure that outputs the graphical representation $G_{\mathscr{D}(\mathbb{P})}$ in (A) given $\mathbb{P}$ and one DAG $D \in \mathscr{D}(\mathbb{P})$. The proposed algorithm only relies on sequences of local transformations and score computations and hence is feasible for large graphs with numbers of variables in the thousands (assuming reasonable sparsity). We demonstrate its performance on simulated data and derive its (high-dimensional) consistency based on the consistency proof of the CAM methodology in Section 2.4.

3.2 Comprehensive characterization of the identifiability of PLSEMs

This section contains our main theoretical results. They consist of characterizations of PLSEMs that generate a given distribution $\mathbb{P}$ from various perspectives. In Section 3.2.1 we assume that $\mathbb{P}$ is faithful to the underlying causal model and demonstrate that this leads to a transformational characterization and a graphical representation of $\mathscr{D}(\mathbb{P})$ very similar to the well-known counterparts characterizing a Markov equivalence class. Our main theoretical contributions, which hold under very general assumptions and, in particular, do not rely on the faithfulness assumption, are presented in Section 3.2.2. They fully characterize all PLSEMs that generate a given distribution $\mathbb{P}$ on a functional level. Moreover, they explain how nonlinear functions impose very specific restrictions on the set of potential causal orderings. Section 3.2.3 brings together the two previous sections by discussing the precise interplay of nonlinearity and faithfulness.

3.2.1 Characterizations of $\mathscr{D}(\mathbb{P})$ under faithfulness

Let $\mathbb{P}$ be generated by a PLSEM with DAG $D \in \mathscr{D}(\mathbb{P})$. The goal of this section is to characterize $\mathscr{D}(\mathbb{P})$. Recall that $\mathscr{D}(\mathbb{P})$ is the set of all DAGs D such that $\mathbb{P}$ is faithful to D and there exists a PLSEM with DAG D that generates $\mathbb{P}$. In words, faithfulness means that no conditional independence relations other than those entailed by the Markov property hold, see, e.g., Spirtes et al. (2000). In particular, it implies that $\mathscr{D}(\mathbb{P})$ is a subset of the Markov equivalence class and all DAGs in $\mathscr{D}(\mathbb{P})$ have the same skeleton and v-structures (Verma and Pearl, 1990). Markov equivalence classes can be graphically represented with single graphs, known as CPDAGs (also referred to as essential graphs, maximally oriented graphs or completed patterns) (Andersson et al., 1997; Chickering, 1995; Meek, 1995; Verma and Pearl, 1990), where an edge is directed if and only if it is oriented the same way in all the DAGs in the Markov equivalence class, else, it is undirected. The Markov equivalence class then equals the set of all DAGs that can be obtained from the CPDAG by orienting the undirected edges without creating new v-structures. We derive an analogous graphical representation of $\mathscr{D}(\mathbb{P})$.

Another useful (transformational) characterization result says that any two Markov equivalent DAGs can be transformed into each other by a

sequence of distinct covered edge reversals (Chickering, 1995). We will demonstrate that a very similar principle transfers to $\mathscr{D}(\mathbb{P})$.

Graphical representation of $\mathscr{D}(\mathbb{P})$

The distribution equivalence class $\mathscr{D}(\mathbb{P})$ can be graphically represented by a single partially directed acyclic graph (PDAG). A PDAG is a graph with directed and undirected edges that does not contain any directed cycles. A *consistent DAG extension* of a PDAG is a DAG with the same skeleton, the same edge orientations on the directed subgraph of the PDAG, and no additional v-structures.

Definition 1. *Let $\mathcal{E}$ be a set of Markov equivalent DAGs. We denote by $G_{\mathcal{E}}$ the PDAG that has the same skeleton as the DAGs in $\mathcal{E}$ and $i \to j$ in $G_{\mathcal{E}}$ if and only if $i \to j$ in all the DAGs in $\mathcal{E}$, else, $i - j$. We say that $G_{\mathcal{E}}$ is maximally oriented with respect to $\mathcal{E}$.*

For a given distribution equivalence class $\mathscr{D}(\mathbb{P})$, the corresponding PDAG $G_{\mathscr{D}(\mathbb{P})}$ is uniquely defined by Definition 1. Moreover, $G_{\mathscr{D}(\mathbb{P})}$ is a graphical representation of $\mathscr{D}(\mathbb{P})$ in the following sense:

Theorem 4. *$\mathscr{D}(\mathbb{P})$ is equal to the set of all consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$.*

A proof can be found in Appendix 3.A.1. Theorem 4 states that one can represent $\mathscr{D}(\mathbb{P})$ with a single PDAG $G_{\mathscr{D}(\mathbb{P})}$ without loss of information, as $\mathscr{D}(\mathbb{P})$ can be reconstructed from $G_{\mathscr{D}(\mathbb{P})}$ by listing all consistent DAG extensions. An example is given in Figure 3.2.

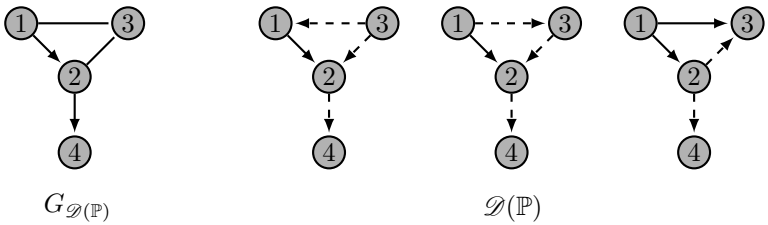


Figure 3.2: Graphical representation of $\mathscr{D}(\mathbb{P})$ with the single PDAG $G_{\mathscr{D}(\mathbb{P})}$. $\mathscr{D}(\mathbb{P})$ equals the set of all consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$. The graph with $2 \to 3 \to 1$ is not a consistent DAG extension of $G_{\mathscr{D}(\mathbb{P})}$ as it contains a cycle. Linear edges are dashed, nonlinear edges are solid.

Note that $G_{\mathscr{D}(\mathbb{P})}$ can be interpreted as a maximally oriented graph with respect to some background knowledge as defined in Meek (1995). For details, we refer to Section 3.3.2.

Conceptually, this result is completely analogous to the use of CPDAGs to represent Markov equivalence classes. There are important differences, though: first of all, necessary and sufficient conditions have been derived for a graph to be a CPDAG of a Markov equivalence class (Andersson et al., 1997, Theorem 4.1). These properties do not all transfer to $G_{\mathscr{D}(\mathbb{P})}$. For example, $G_{\mathscr{D}(\mathbb{P})}$ typically is not a chain graph, see Figure 3.2. Secondly, given a DAG D, the CPDAG (and hence a full characterization of the Markov equivalence class) can be obtained by an iterative application of three purely graphical orientation rules (R1-R3 in Figure 3.6) applied to the pattern of D (Meek, 1995). This is not true for $G_{\mathscr{D}(\mathbb{P})}$ and $\mathscr{D}(\mathbb{P})$. It is still feasible to obtain $G_{\mathscr{D}(\mathbb{P})}$ from a DAG $D \in \mathscr{D}(\mathbb{P})$, but it is crucial to know which of the functions in the (unique) corresponding PLSEM (cf. Lemma 8) are linear and which are nonlinear. We will show in Section 3.3 that the transformational characterization in Theorem 5 gives rise to a consistent and efficient score-based procedure to estimate $G_{\mathscr{D}(\mathbb{P})}$ based on $D \in \mathscr{D}(\mathbb{P})$ and samples of $\mathbb{P}$.

Transformational characterization of $\mathscr{D}(\mathbb{P})$

Given $D \in \mathscr{D}(\mathbb{P})$, the distribution equivalence class $\mathscr{D}(\mathbb{P})$ can be comprehensively characterized via sequences of local transformations of DAGs.

Theorem 5. *Assume that $\mathbb{P}$ has been generated by a PLSEM and that it is faithful to the underlying DAG. Then, the following two results hold:*

(a) Let $D \in \mathscr{D}(\mathbb{P})$, $i \to j$ covered in D, and D' be the DAG that differs from D only by the reversal of $i \to j$. Then, $D' \in \mathscr{D}(\mathbb{P})$ if and only if $i \to j$ is linear in D. Furthermore, if $i \to j$ is covered and nonlinear in D, then $i \to j$ in all DAGs in $\mathscr{D}(\mathbb{P})$.

(b) Let $D, D' \in \mathscr{D}(\mathbb{P})$. Then there exists a sequence of distinct covered linear edge reversals that transforms D to D'.

A proof can be found in Appendix 3.A.2 and an illustration is provided in Figure 3.3. Note that the interesting part of this result is that $\mathscr{D}(\mathbb{P})$ is connected with respect to covered linear edge reversals. It will be of

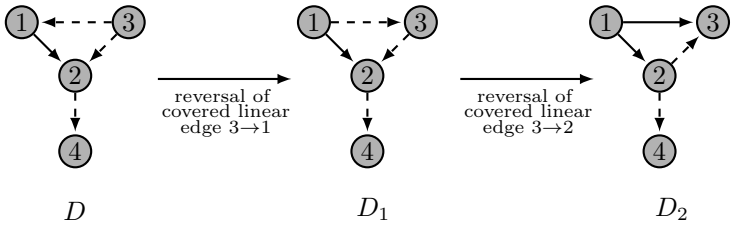


Figure 3.3: Transformational characterization of $\mathscr{D}(\mathbb{P})$ from Figure 3.2. Let $1 \to 2$ in D be nonlinear (solid) and all other edges in D be linear (dashed). Then, D_1 and D_2 can be reached from D by the displayed sequence of covered linear edge reversals. Note that in D and D_2, $1 \to 2$ is covered but nonlinear and hence cannot be reversed by Theorem 5 (a). Moreover, $2 \to 4$ is not covered in any of D, D_1 and D_2 and hence cannot be reversed.

particular importance in the design of score-based estimation procedures for $\mathscr{D}(\mathbb{P})$ and $G_{\mathscr{D}(\mathbb{P})}$ in Section 3.3.

Theorem 5 covers the two special cases discussed in Section 3.1.2: if all the functions $f_{j,i}$ in equation (3.2) are linear, $\mathscr{D}(\mathbb{P})$ (which, in this setting, is equal to the Markov equivalence class) can be fully characterized by sequences of covered edge reversals of D (as all the edges are linear). If, on the contrary, all the functions $f_{j,i}$ in equation (3.2) are nonlinear, $\mathscr{D}(\mathbb{P})$ only consists of the DAG D as there is no covered linear edge in D.

3.2.2 Characterizations not assuming faithfulness

In this section we give general characterizations of PLSEMs that generate the same distribution $\mathbb{P}$, both, from a functional viewpoint and from the perspective of causal orderings. The former describes how the $f_{j,i}$ of different PLSEMs relate to each other, the latter describes the set of causal orderings, such that there exists a corresponding PLSEM that generates the given distribution $\mathbb{P}$. It will show that nonlinear functions impose a very specific structure on the model, which (perhaps surprisingly) is compatible with some of the previous theory on graphical models, as described in Section 3.1.2. Furthermore it will help us understand in the general case how nonlinear functions restrict the set of PLSEMs that generate $\mathbb{P}$. Lastly, we give some additional intuition on the functional characterization. Throughout this section, we assume that $\mathbb{P}$ is generated by a PLSEM as defined in equation (3.2).

Functional characterization

Let us first characterize the result on the level of SEMs. Consider a PLSEM that generates $\mathbb{P}$,

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) + \varepsilon_j,$$

where $f_{j,i}, D, \varepsilon_j, \mu_j$ and $\sigma_j^2 = \mathrm{Var}(\varepsilon_j)$ satisfy the assumptions from Section 3.1.1.

Let us define the function $F : \mathbb{R}^p \to \mathbb{R}^p$ by

$$F(x)_j := \frac{1}{\sigma_j} \left(x_j - \mu_j - \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(x_i) \right). \tag{3.3}$$

It turns out to be convenient to work with this function F. Notably, we do not lose any information by working with F instead of $f_{j,i}$, $\mathrm{pa}_D(j), \mu_j$ and σ_j as these quantities can be recovered from F. Specifically, we can easily obtain the distribution of the errors from the function F as

$$\sigma_j := 1/\partial_j F_j. \tag{3.4}$$

By definition, $F(X) \sim \mathcal{N}(0, \mathrm{Id}_p)$. Hence, for $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$ it holds that $F^{-1}(Z) \sim X$. Using this, we obtain $\mu_j = \mathbb{E}_Z[F^{-1}(Z)_j]$ and we can recover the functions $f_{j,i}$ from the function F using the equations

$$f'_{j,i} = -\sigma_j \partial_i F_j \qquad \text{and} \qquad \mathbb{E}_Z f_{j,i}(F^{-1}(Z)_i) = 0. \tag{3.5}$$

Note that the equation on the left hand side determines $f_{j,i}$ up to a constant, whereas the equation on the right hand side determines the constant using only quantities that can be calculated from F. In the same spirit, $\mathrm{pa}_D(j)$ can be recovered from F via

$$\mathrm{pa}_D(j) = \{i \neq j : \partial_i F_j \not\equiv 0\}. \tag{3.6}$$

In this sense, instead of describing the PLSEM by $f_{j,i}, \mathrm{pa}_D(j), \mu_j$ and σ_j it can simply be described by the function $F : \mathbb{R}^p \to \mathbb{R}^p$. Now let us define

$$\mathcal{F}(\mathbb{P}) := \{F \colon \mathbb{R}^p \mapsto \mathbb{R}^p \mid F \text{ suffices (3.3) for a PLSEM that generates } \mathbb{P}\}.$$

We call the functions in this set *PLSEM-functions*. Let us define the set of orthonormal matrices $\mathcal{O}_n(\mathbb{R}) = \{O \in \mathbb{R}^{p \times p} : OO^t = \mathrm{Id}\}$. The following theorem follows from Theorem 10 that is given in Appendix 3.A.3. It states that we can construct all PLSEMs that generate $\mathbb{P}$ by essentially *rotating* F.

Theorem 6 (Characterization of potential PLSEMs)**.** *For $F \in \mathcal{F}(\mathbb{P})$ there exists a set of (constant) rotations $\mathcal{O}_{\mathcal{F}(\mathbb{P})} \subset \mathcal{O}_n(\mathbb{R})$ such that*

$$\mathcal{F}(\mathbb{P}) = \left\{O \cdot F : O \in \mathcal{O}_{\mathcal{F}(\mathbb{P})}\right\}.$$

A description and explicit formulae for each $O \in \mathcal{O}_{\mathcal{F}(\mathbb{P})}$ are given in Remark 8 in Appendix 3.A.3.

Astonishingly, in this sense, all PLSEMs that generate $\mathbb{P}$ are rotations of each other. The importance of this result lies in its simplicity: There are very simple linear relationships between the $f_{j,i}$ in one PLSEM and the $\tilde{f}_{j,i}$ in another PLSEM. The formulae in Appendix 3.A.3 permit to fully characterize these matrices $\mathcal{O}_{\mathcal{F}(\mathbb{P})}$. In fact, the characterization in Theorem 10 is the first step towards all other characterizations.

Intuition on the functional characterization

This section motivates Theorem 6. Consider two functions $F, G \in \mathcal{F}(\mathbb{P})$ that correspond to two different PLSEMs that generate $\mathbb{P}$. By Proposition 1 in Appendix 3.A.3,

$$F(X) \sim \mathcal{N}(0, \mathrm{Id}) \text{ and } G(X) \sim \mathcal{N}(0, \mathrm{Id}). \tag{3.7}$$

Moreover, it follows from the definition of PLSEMs that F is invertible. Let $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$. Using equation (3.7) twice,

$$F^{-1}(Z) \sim X \text{ and } G(F^{-1}(Z)) \sim \mathcal{N}(0, \mathrm{Id}).$$

Hence $J : \mathbb{R}^p \to \mathbb{R}^p$, $J := G \circ F^{-1}$ suffices $J(Z) \sim Z \sim \mathcal{N}(0, \mathrm{Id})$. Furthermore, it can be shown that $|\det \mathrm{D}J| = 1$. Using the transformation formula, we obtain

$$\frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|J(x)\|_2^2}{2}\right) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|x\|_2^2}{2}\right) \text{ for all } x \in \mathbb{R}^p.$$

By rearranging,

$$\|J(x)\|_2 = \|x\|_2 \text{ for all } x \in \mathbb{R}^p.$$

If we admit that J must be a linear function (which requires some work), this formula gives us $J \in \mathcal{O}_n(\mathbb{R}) := \{O \in \mathbb{R}^{p \times p} : OO^t = \mathrm{Id}\}$ and it immediately follows that $G = JF$. This reasoning shows that the main work in proving Theorem 6 lies in showing that J is a linear function.

Characterization via causal orderings

This section discusses a characterization of all potential causal orderings of a given PLSEM. Let us define the set of *potential causal orderings* as

$$\mathcal{S}(\mathbb{P}) := \left\{ \begin{array}{l} \sigma \text{ permutation on } \{1,...,p\} : \text{there exists} \\ \text{a PLSEM with DAG } D \text{ that generates } \mathbb{P} \\ \text{such that } \sigma(i) < \sigma(j) \text{ for all } i \to j \text{ in } D \end{array} \right\}.$$

Without assuming faithfulness, if all $f_{j,i}$ are linear, all permutations of $\{1,...,p\}$ are a causal ordering of a DAG corresponding to a PLSEM that generates $\mathbb{P}$. That is, $\mathcal{S}(\mathbb{P})$ is equal to the set of all permutations of $\{1,\dots,p\}$. Roughly, the more nonlinear functions in the PLSEM, the smaller the resulting set $\mathcal{S}(\mathbb{P})$. The interesting point is that nonlinear edges restrict $\mathcal{S}(\mathbb{P})$ in a very specific way. Before we state the theorem, consider a PLSEM that generates $\mathbb{P}$, define the function $F : \mathbb{R}^p \to \mathbb{R}^p$ as in equation (3.3) and define the set

$$\mathcal{V} := \{(i,j) \in \{1,...,p\}^2 : e_j^t (\mathrm{D}F)^{-1} \partial_i^2 F \not\equiv 0\}, \tag{3.8}$$

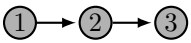
where e_j, $j = 1,\dots,p$ denotes the standard basis of $\mathbb{R}^p$ and $\mathrm{D}F$ is the Jacobian of F. In some sense, we can think of $e_j^t (\mathrm{D}F)^{-1} \partial_i^2 F \not\equiv 0$ as "the effect from variable i to variable j is nonlinear". We will discuss the set $\mathcal{V}$ in more detail later. The potential causal orderings can now be characterized as follows:

Theorem 7 (Characterization of potential causal orderings)**.**

$$\mathcal{S}(\mathbb{P}) = \{\sigma \textit{ permutation on } \{1,\dots,p\} : \sigma(i) < \sigma(j) \textit{ for all } (i,j) \in \mathcal{V}\}$$

A proof of this theorem can be found in Appendix 3.A.4. In words, all permutations of the indices that do not swap any of the tuples in $\mathcal{V}$ are a causal ordering of a DAG corresponding to a PLSEM that generates $\mathbb{P}$. And for all permutations of indices for which one of the tuples in $\mathcal{V}$ is switched, there exists *no* PLSEM with this causal ordering that generates $\mathbb{P}$. Moreover, by Theorem 11 (b), if $(i,j) \in \mathcal{V}$, then j is a descendant of i in every PLSEM that generates $\mathbb{P}$. Now let us give some intuition on the index tuples in the set $\mathcal{V}$.

Example 1. *Consider the DAG*



and the distribution $\mathbb{P}$ *that has been generated by a PLSEM of the form*

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2 &= f_{2,1}(X_1) + \varepsilon_2 \\ X_3 &= f_{3,2}(X_2) + \varepsilon_3, \end{aligned}$$

with $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_3)$.

(a) Let $f_{2,1}(x) = 0.5x$ *be linear,* $f_{3,2}(x) = x^3$ *be nonlinear. The corresponding PLSEM-function is* $F(x) = (x_1, x_2 - 0.5x_1, x_3 - x_2^3)^t$. *Using elementary calculations it can be seen that* $e_j^t(\mathrm{D}F)^{-1}\partial_i^2 F \not\equiv 0$ *only for* $(i,j) = (2,3)$. *Hence,* $\mathcal{V} = \{(2,3)\}$ *and all permutations* σ *respecting* $\sigma(2) < \sigma(3)$ *are a causal ordering of a DAG corresponding to a PLSEM that generates* $\mathbb{P}$. *For example, for the causal ordering* $\sigma(2) < \sigma(3) < \sigma(1)$, *there exists a (unique) PLSEM with DAG* $1 \leftarrow 2 \rightarrow 3$ *that generates* $\mathbb{P}$.

(b) Let $f_{2,1}(x) = x^3$ *be nonlinear,* $f_{3,2}(x) = 0.5x$ *be linear. The corresponding PLSEM-function is* $F(x) = (x_1, x_2 - x_1^3, x_3 - 0.5x_2)^t$. *We obtain* $\mathcal{V} = \{(1,2),(1,3)\}$ *and all permutations* σ *with* $\sigma(1) < \sigma(2)$ *and* $\sigma(1) < \sigma(3)$ *are a causal ordering of a DAG corresponding to a PLSEM that generates* $\mathbb{P}$. *In particular, for* $\sigma(1) < \sigma(3) < \sigma(2)$ *we obtain that the PLSEM corresponding to the (unfaithful) DAG* $1 \rightarrow 3 \rightarrow 2$ *with* $1 \rightarrow 2$ *generates* $\mathbb{P}$.

Let us make several concluding remarks: in (a), the causal ordering between nodes 1 and 3 is not fixed, whereas in (b), it is fixed. Hence, the set $\mathcal{V}$ sometimes also fixes the causal ordering between two nodes that are not adjacent in the DAG corresponding to F. Secondly, in both examples, the causal ordering of nodes incident to nonlinear edges is fixed. This raises the question whether it is true in general that nonlinear edges cannot be reversed. The answer is no (see Figure 3.4), but in some sense, the models with "reversible nonlinear edges" are rather particular. Finally, if we make additional mild assumptions, stronger statements can be made about the index tuples in $\mathcal{V}$. This will be discussed in the next section.

3.2.3 The interplay of nonlinearity and faithfulness

As indicated in Section 3.2.2, without further assumptions, some nonlinear edges can be reversed. An example is given in Figure 3.4.

Figure 3.4: Nonlinear edges can be reversed if nonlinear effects cancel out. $X_1 = \varepsilon_1$, $X_2 = X_1^2 + X_1 + \varepsilon_2$, $X_3 = X_2 - X_1^2 + \varepsilon_3$ with $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_3)$ generates the same joint distribution of (X_1, X_2, X_3) as $X_3 = \tilde{\varepsilon}_3$, $X_1 = X_3/3 + \tilde{\varepsilon}_1$, $X_2 = X_1/2 + X_1^2 + X_3/2 + \tilde{\varepsilon}_2$ with $\tilde{\varepsilon}_3 \sim \mathcal{N}(0, 3), \tilde{\varepsilon}_1 \sim \mathcal{N}(0, 2/3), \tilde{\varepsilon}_2 \sim \mathcal{N}(0, 1/2)$ independent. This stems from the fact that the nonlinear parts of the functions $f_{2,1}(x)$ and $f_{3,1}(x)$ cancel out, that is, $f''_{2,1} + f''_{3,1} = 0$. Note that this example does not contradict the previous theoretical results. It holds that $e_3^t(\mathrm{D}F)^{-1}\partial_1^2 F \equiv 0$ for the PLSEM-function F corresponding to D_1. Hence the causal ordering of D_2 does not contradict Theorem 7.

The edge $1 \to 3$ in D_1 can be reversed even though $f_{3,1}$ is a nonlinear function in the PLSEM corresponding to D_1. This issue arises because the nonlinear effect from X_1 to X_3 in D_1 cancels out over two paths. If we write X_3 as a function of $\varepsilon_1, \varepsilon_2, \varepsilon_3$, that function is linear. The setting of D_1 in Figure 3.4 is rather particular as $\partial_1^2 f_{2,1}$ and $\partial_1^2 f_{3,1}$ are linearly dependent. As the function space $\mathcal{C}^2(\mathbb{R})$ is infinite dimensional, this is arguably a degenerate scenario. Note that faithfulness does not save us from this cancellation effect as $\mathbb{P}$ is faithful to both, D_1 and D_2.

Nevertheless, we can rely on a different, rather weak assumption: consider a node i in a DAG D and assume that the corresponding functions in the set

$$\{\partial_i^2 f_{j',i} : j' \text{ is a child of } i \text{ in } D \text{ and } f_{j',i} \text{ is nonlinear}\}$$

are linearly independent. In other words: assume that the "nonlinear effects" from X_i on its children are linearly independent functions. Then these nonlinear edges cannot be reversed.

The following corollary is a direct implication of Theorem 11 (a) and (b) in Appendix 3.A.5.

Corollary 1. *Consider a PLSEM and the corresponding distribution $\mathbb{P}$. Let j be a child of i in D and let $f_{j,i}$ be a nonlinear function. If the functions in the set $\{\partial_i^2 f_{j',i} : j'$ is a child of i in D and $f_{j',i}$ is nonlinear$\}$ are linearly independent, then j is a descendant of i in any other DAG D' of a PLSEM that generates $\mathbb{P}$.*

Intuitively, this should not be the end of the story: if an edge $i \to j$ is nonlinear, then usually there should also be a nonlinear relationship

between i and the descendants of j. Hence it should be possible to infer some statements about the causal ordering of i and the descendants of j. In general, this is not true as demonstrated in Figure 3.5.

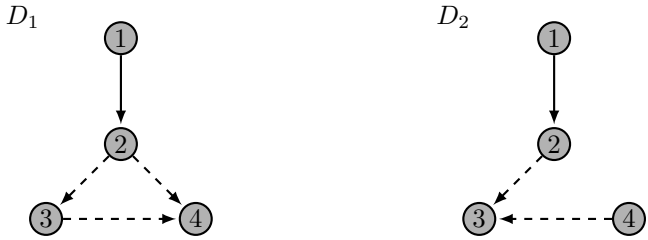


Figure 3.5: If $\mathbb{P}$ is not faithful to D, descendants are not fixed. Node 4 is a descendant of node 1 in D_1 but not in D_2. On the left hand side, $X_1 = \varepsilon_1$, $X_2 = X_1^2 + \varepsilon_2$, $X_3 = X_2 + \varepsilon_3$, $X_4 = X_3 - X_2 + \varepsilon_4$, with $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_4)$. On the right hand side, $X_1 = \tilde{\varepsilon}_1$, $X_2 = X_1^2 + \tilde{\varepsilon}_2$, $X_3 = X_2 + 1/2 \cdot X_4 + \tilde{\varepsilon}_3$, $X_4 = \tilde{\varepsilon}_4$, where $\tilde{\varepsilon}_1 \sim \mathcal{N}(0,1), \tilde{\varepsilon}_2 \sim \mathcal{N}(0,1), \tilde{\varepsilon}_3 \sim \mathcal{N}(0, 1/2)$ and $\tilde{\varepsilon}_4 \sim \mathcal{N}(0, 2)$. Both PLSEMs generate the same distribution. Note that in this case, additional assumptions on the nonlinear function $f_{2,1}$ would not resolve the issue.

Under the assumption of faithfulness, additional statements can be made about descendants of j. In some sense, the nonlinear effect from i on the descendants of j, mediated through some of the descendants of j, cannot "cancel out". Hence, all descendants of j are fixed. The following corollary is a direct implication of Theorem 11 (c) and (d) in Appendix 3.A.5.

Corollary 2. *Let the assumptions of Corollary 1 be true. In addition, let $\mathbb{P}$ be faithful to the DAG D. Fix $k \neq i$. Then k is a descendant of i in each DAG D' of a PLSEM that generates $\mathbb{P}$ if and only if k is a descendant of a nonlinear child of i in D.*

Note that we use the convention that a node is a descendant of itself. Corollary 2 guarantees that certain descendants of i are descendants of i in all DAGs D' of PLSEMs that generate $\mathbb{P}$. In that sense, it provides a simple criterion that tells us whether or not k is descendant of i in all of these DAGs. It is crucial to be precise: we do not assume that $\mathbb{P}$ is faithful to D', that means, we search over all PLSEMs that generate $\mathbb{P}$. If we search over the smaller space $\mathscr{D}(\mathbb{P})$, that is, additionally assume that $\mathbb{P}$ is faithful to D', the set of potential PLSEMs usually gets smaller. In many cases, there are some edges that are not fixed if we search over all PLSEMs, but fixed if we only search over PLSEMs with DAGs in $\mathscr{D}(\mathbb{P})$.

As discussed in Section 3.2.1, $\mathscr{D}(\mathbb{P})$ can be represented by a single PDAG $G_{\mathscr{D}(\mathbb{P})}$. In the following, we will discuss the estimation of $\mathscr{D}(\mathbb{P})$ and $G_{\mathscr{D}(\mathbb{P})}$.

3.3 Score-based estimation

Consider $\mathbb{P}$ that has been generated by a PLSEM and assume that $\mathbb{P}$ is faithful to the underlying DAG. We denote by $\{X^{(i)}\}_{i=1,\dots,n}$ i.i.d. copies of $X \in \mathbb{R}^p$ and by $\mathbb{P}_n$ their empirical distribution. The goal of this section is to derive a consistent score-based estimation procedure for the distribution equivalence class $\mathscr{D}(\mathbb{P})$ based on $\mathbb{P}_n$ and one (true) DAG $D^0 \in \mathscr{D}(\mathbb{P})$. We first describe a "naive" recursive solution that lists all members of $\mathscr{D}(\mathbb{P})$ and motivate the score-based approach in Section 3.3.1. We then present a more efficient procedure that directly estimates the graphical representation $G_{\mathscr{D}(\mathbb{P})}$ as defined in Section 3.3.2. Both methods rely on the transformational characterization result in Theorem 5.

In practice, we may replace the true D^0 by an estimate, e.g., from the CAM methodology (Bühlmann et al., 2014); see Chapter 2. If the estimate is consistent for a DAG in $\mathscr{D}(\mathbb{P})$ we obtain consistency of our method for the entire distribution equivalence class $\mathscr{D}(\mathbb{P})$.

3.3.1 Estimation of $\mathscr{D}(\mathbb{P})$

Theorem 5 provides a straightforward way to list all members of $\mathscr{D}(\mathbb{P})$. Starting from the DAG D^0, one can search over all sequences of distinct covered linear edges reversals. By Theorem 5 (a), all DAGs that are traversed are in $\mathscr{D}(\mathbb{P})$ and by Theorem 5 (b), $\mathscr{D}(\mathbb{P})$ is connected with respect to sequences of distinct covered linear edge reversals. Moreover, by Theorem 5 (a), an edge that is nonlinear and covered in a DAG in $\mathscr{D}(\mathbb{P})$ has the same orientation in all the members of $\mathscr{D}(\mathbb{P})$. These simple observations immediately lead to a recursive estimation procedure. Its population version is described in Algorithm 1. The inputs are D^0 (with all its edges marked as "unfixed") and an oracle that answers the question if a specific edge in a DAG in $\mathscr{D}(\mathbb{P})$ is linear or nonlinear.

Unfortunately, the (true) information if a selected covered edge $i \to j$ in a DAG $D \in \mathscr{D}(\mathbb{P})$ is linear or not is generally not available. Also, it cannot simply be deduced from the starting DAG D^0 as the status of the edge may have changed in D. For an example, see Figure 3.1: edge $1 \to 3$ is

Algorithm 1 listAllDAGsPLSEM (population version)

if there is no covered edge in DAG D^0 that is marked as unfixed **then**
  Add D^0 to the distribution equivalence class $\mathscr{D}(\mathbb{P})$ and terminate.
end if
Choose a covered edge $i \to j$ in DAG D^0 that is marked as unfixed.
if the edge $i \to j$ is linear in D^0 **then**
  Define a DAG $D_1^0 := D^0$ with edge $i \to j$ in D_1^0 marked as fixed and a DAG D_2^0 equal to D^0 except for a reversed edge $i \leftarrow j$ marked as fixed in D_2^0.
  Call `listAllDAGsPLSEM` recursively for both DAGs D_1^0 and D_2^0.
else
  Mark $i \to j$ in D^0 as fixed and call `listAllDAGsPLSEM` for DAG D^0.
end if

not covered and linear in D_1 but nonlinear and covered in $D_2 \in \mathscr{D}(\mathbb{P})$, and hence irreversible.

To check the status of a covered edge in a given DAG $D \in \mathscr{D}(\mathbb{P})$, one could either test (non-)linearity of the functional component in the (unique) PLSEM corresponding to D or rely on a score-based approach. In the following we are going to elaborate on the latter. We closely follow the approach presented in Bühlmann et al. (2014), see also Section 2.2.

We assume that the functions $f_{j,i}$ in equation (3.2) are from a class of smooth functions $\mathcal{F}_i \subseteq \{f \in C^2(\mathbb{R}), \mathbb{E}[f(X_i)] = 0\}$, which is closed with respect to the $L_2(\mathbb{P}_{X_i})$-norm and closed under linear transformations. For a set of given basis functions, we denote by $\mathcal{F}_{n,i} \subseteq \mathcal{F}_i$ the finite-dimensional approximation space which typically increases as n increases. The spaces of additive functions with components in $\mathcal{F}_i$ and $\mathcal{F}_{n,i}$, respectively, are closed assuming an analogue of a minimal eigenvalue condition. All details are given in Bühlmann et al. (2014) or in Section 2.2. Without loss of generality, we assume $\mu_j = 0$ as in the original paper. For $D^0 \in \mathscr{D}(\mathbb{P})$, let $\theta^{D^0} := (\{f_{j,i}^{D^0}\}_{j=1,\dots,p, i \in \mathrm{pa}_{D^0}(j)}, \{\sigma_j^{D^0}\}_{j=1,\dots,p})$ be the infinite-dimensional parameter of the corresponding PLSEM. The expected negative log-likelihood reads

$$\mathbb{E}\left[-\log p_{\theta^{D^0}}(X)\right] = \sum_{j=1}^{p} \log(\sigma_j^{D^0}) + \mathrm{C}, \qquad C = \frac{p}{2}\log(2\pi) + \frac{p}{2}.$$

All $D^0 \in \mathscr{D}(\mathbb{P})$ lead to the minimal expected negative log-likelihood, as

by definition, the corresponding PLSEM generates the true distribution $\mathbb{P}$. For a misspecified model with wrong DAG $D \notin \mathscr{D}(\mathbb{P})$ we obtain the projected parameter $\theta^D = \left(\{f^D_{j,i}\}_{j=1,\dots,p, i\in \mathrm{pa}_D(j)}, \{\sigma^D_j\}_{j=1,\dots,p}\right)$ as

$$\begin{aligned}\{f^D_{j,i}\}_{i\in\mathrm{pa}_D(j)} &= \operatorname*{argmin}_{g_{j,i}\in\mathcal{F}_i} \mathbb{E}[(X_j - \sum_{i\in\mathrm{pa}_D(j)} g_{j,i}(X_i))^2] \\ (\sigma^D_j)^2 &= \mathbb{E}[(X_j - \sum_{i\in\mathrm{pa}_D(j)} f^D_{j,i}(X_i))^2]\end{aligned}$$

with expected negative log-likelihood

$$\mathbb{E}\left[-\log\left(p^D_{\theta^D}(X)\right)\right] = \sum_{j=1}^{p} \log(\sigma^D_j) + C, \qquad C = \frac{p}{2}\log(2\pi) + \frac{p}{2},$$

where all expectations are taken with respect to the true distribution $\mathbb{P}$. We refer to $\mathbb{E}\left[-\log\left(p^D_{\theta^D}(X)\right)\right]$ as the *score of* D and to $\log(\sigma^D_j)$ as *score* of node j in D. For a DAG $D^0 \in \mathscr{D}(\mathbb{P})$, let

$$\mathscr{C}(D^0) = \left\{ D \;\middle|\; \begin{array}{l} D \text{ and } D^0 \text{ differ by a single} \\ \text{covered nonlinear edge reversal} \end{array} \right\}.$$

Then, for $D^0 \in \mathscr{D}(\mathbb{P})$ and $D \in \mathscr{C}(D^0)$ that (without loss of generality) only differ by the orientation of the covered edge between the nodes i and j, the difference in expected negative log-likelihood is given as

$$\begin{aligned}&\mathbb{E}\left[-\log\left(p^D_{\theta^D}(X)\right)\right] - \mathbb{E}\left[-\log\left(p_{\theta^{D^0}}(X)\right)\right] \\ &= \log(\sigma^D_i) + \log(\sigma^D_j) - \log(\sigma^{D^0}_i) - \log(\sigma^{D^0}_j).\end{aligned} \tag{3.9}$$

Since the score is decomposable over the nodes, the reversal of a covered edge only affects the scores locally at the two nodes i and j incident to the covered edge. We denote by

$$\xi_p := \min_{\substack{D^0\in\mathscr{D}(\mathbb{P}) \\ D\in\mathscr{C}(D^0)}} \left(\mathbb{E}\left[-\log\left(p^D_{\theta^D}(X)\right)\right] - \mathbb{E}\left[-\log\left(p_{\theta^{D^0}}(X)\right)\right]\right) \tag{3.10}$$

the *degree of separation* of true models in $\mathscr{D}(\mathbb{P})$ and misspecified models in $\mathscr{C}(\mathscr{D}(\mathbb{P}))$ that can be reached by the reversal of one covered nonlinear edge in any DAG $D^0 \in \mathscr{D}(\mathbb{P})$. From the transformational characterization in Theorem 5 it follows that $\xi_p > 0$. Combining equations (3.9) and (3.10) motivates the estimation procedure in Algorithm 2 that takes as inputs n

samples $X^{(1)}, ..., X^{(n)}$ and a DAG $D^0 \in \mathscr{D}(\mathbb{P})$ (with all its edges marked as "unfixed") and outputs a score-based estimate $\widehat{\mathscr{D}}_{n,p}$ of $\mathscr{D}(\mathbb{P})$.

Algorithm 2 listAllDAGsPLSEM

if there is no covered edge in DAG D^0 that is marked as unfixed **then**
  Add D^0 to $\widehat{\mathscr{D}}_{n,p}$ and terminate.
end if
Choose a covered edge $i \to j$ in DAG D^0 that is marked as unfixed. Let D' be the DAG that equals D^0 except for a reversed edge $i \leftarrow j$.
Additively regress X_i on $X_{\mathrm{pa}_{D^0}(i)}$, X_j on $X_{\mathrm{pa}_{D^0}(j)}$, X_i on $X_{\mathrm{pa}_{D^0}(i)\cup\{j\}}$, X_j on $X_{\mathrm{pa}_{D^0}(i)}$
Compute the standard deviations of the residuals to obtain $\hat{\sigma}_i^{D^0}$, $\hat{\sigma}_j^{D^0}$, $\hat{\sigma}_i^{D'}$ and $\hat{\sigma}_j^{D'}$.
if $|\log(\hat{\sigma}_i^{D'}) + \log(\hat{\sigma}_j^{D'}) - \log(\hat{\sigma}_i^{D^0}) - \log(\hat{\sigma}_j^{D^0})| < \alpha$ **then**
  Set $D_1^0 := D^0$ with $i \to j$ marked as fixed and $D_2^0 := D'$ with $i \leftarrow j$ marked as fixed.
  Call `listAllDAGsPLSEM` recursively for both DAGs D_1^0 and D_2^0.
else
  Mark $i \to j$ in D^0 as fixed and call `listAllDAGsPLSEM` for DAG D^0.
end if

To prove the (high-dimensional) consistency of the score-based estimation procedure, we make the following assumptions. For a function $h : \mathbb{R} \to \mathbb{R}$, we write $P(h) = \mathbb{E}[h(X)]$ and $P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X^{(i)})$.

Assumption 1.

(i) Uniform upper bound on node degrees:

$$\max_{\substack{D\in\mathscr{D}(\mathbb{P})\cup\mathscr{C}(\mathscr{D}(\mathbb{P}))\\ j=1,\dots,p}} \deg_D(j) \le M \text{ for some positive constant } M < \infty.$$

(ii) Uniform lower bound on error variances:

$$\min_{\substack{D\in\mathscr{D}(\mathbb{P})\cup\mathscr{C}(\mathscr{D}(\mathbb{P}))\\ j=1,\dots,p}} (\sigma_j^D)^2 \ge L > 0.$$

(iii) Empirical process bound:

$$\max_{\substack{D\in\mathscr{D}(\mathbb{P})\cup\mathscr{C}(\mathscr{D}(\mathbb{P}))\\ j=1,\dots,p}} \Delta_{n,j}^{D} = o_P(1),$$

where $\Delta_{n,j}^{D} = \sup_{g_{j,i}\in\mathcal{F}_i} |(P_n - P)((X_j - \sum_{i\in \mathrm{pa}_D(j)} g_{j,i}(X_i))^2)|$.

(iv) Control of approximation error:

$$\max_{\substack{D^0\in\mathscr{D}(\mathbb{P})\\ j=1,\dots,p}} |\gamma_{n,j}^{D^0}| = o(1),$$

where

$$\gamma_{n,j}^{D^0} = \mathbb{E}[(X_j - \sum_{i\in \mathrm{pa}_{D^0}(j)} f_{n;j,i}^{D^0}(X_i))^2] - \mathbb{E}[(X_j - \sum_{i\in \mathrm{pa}_{D^0}(j)} f_{j,i}^{D^0}(X_i))^2]$$

with

$$f_{n;j,i}^{D^0} = \operatorname*{argmin}_{g_{j,i}\in\mathcal{F}_{n,i}} \mathbb{E}[(X_j - \sum_{i\in \mathrm{pa}_{D^0}(j)} g_{j,i}(X_i))^2]$$

and $\mathcal{F}_{n,i}$ *are the approximation spaces as introduced before.*

Assumption 1 (i) is satisfied if D^0 has bounded node degrees, as all DAGs under consideration are restricted to the same skeleton and hence all have equal node degrees. In the low-dimensional setting, Assumption 1 (iii) is justified by Lemma 5 in the supplement to Bühlmann et al. (2014) under the assumptions mentioned there. In the high-dimensional setting, it follows from Lemma 6 in the supplement to Bühlmann et al. (2014) and $\sqrt{\log(p)/n} = o(1)$ together with Assumption 1 (i) and the assumptions mentioned in the original paper. Assumption 1 (iv) can be ensured by requiring a smoothness condition on the coefficients of the basis expansion for the true functions (Bühlmann et al., 2014, Section 4.2). A proof of Theorem 8 can be found in Appendix 3.A.6.

Theorem 8. *Under Assumption 1 and* $\xi_p \geq \xi_0 > 0$*, for any* $\alpha \in (0, \xi_0)$*,*

$$\mathbb{P}[\widehat{\mathscr{D}}_{n,p} = \mathscr{D}(\mathbb{P})] \to 1 \qquad (n \to \infty)$$

Remark 4. *The assumption on the degree of separation of true and wrong models in Bühlmann et al. (2014) is stricter and would imply the uniform bound* $\xi_p/p \geq \xi_0 > 0$*, whereas here we only require* $\xi_p \geq \xi_0 > 0$*. As we are*

given a true DAG $D^0 \in \mathscr{D}(\mathbb{P})$, we solely perform local transformations of DAGs thanks to the transformational characterization result in Theorem 5. This only affects the scores of two nodes and allows us to rely on this much weaker gap condition.

3.3.2 Estimation of $G_{\mathscr{D}(\mathbb{P})}$

The estimation of all DAGs in $\mathscr{D}(\mathbb{P})$ is feasible but may be computationally intractable in the presence of many linear edges. For example, if D^0 is a fully connected DAG with p nodes and all its edges are linear, the number of DAGs in $\mathscr{D}(\mathbb{P})$ corresponds to the number of causal orderings of p nodes which is $p!$. It therefore would be desirable to have a procedure that works without enumerating all DAGs in $\mathscr{D}(\mathbb{P})$. In this section we are going to describe such a procedure that directly estimates the maximally oriented PDAG $G_{\mathscr{D}(\mathbb{P})}$ defined in Section 3.2.1. Recall that by Theorem 4, this fully characterizes $\mathscr{D}(\mathbb{P})$, as $\mathscr{D}(\mathbb{P})$ can be recovered from $G_{\mathscr{D}(\mathbb{P})}$ by listing all consistent DAG extensions.

The main idea is the following: instead of traversing the space of DAGs, we traverse the space of maximally oriented PDAGs that represent sets of distribution equivalent DAGs. As an example, let $D^0 \in \mathscr{D}(\mathbb{P})$ and $i \to j$ be covered and linear in D^0. By Theorem 5 (a), the DAG D' that only differs from D^0 by the reversal of $i \to j$ is in $\mathscr{D}(\mathbb{P})$. Instead of memorizing both, D^0 and D', and recursively searching over sequences of covered linear edge reversals from both of these DAGs as in Algorithms 1 and 2, we represent D^0 and D' by the PDAG G that is maximally oriented with respect to the set of DAGs $\{D^0, D'\}$. By Definition 1, G equals D^0 but for an undirected edge $i - j$. To construct $G_{\mathscr{D}(\mathbb{P})}$, the idea is now to iteratively modify G by either fixing or removing orientations of directed edges if they are nonlinear or linear in one of the consistent DAG extensions of G in which they are covered. For that to work based on G only, that is, without listing all consistent DAG extensions of G, the two key questions are the following:

(Q1) For $i \to j$ in a maximally oriented PDAG G, can we decide based on G only if there is a consistent DAG extension of G in which $i \to j$ is covered?

(Q2) If $i \to j$ is known to be covered in a consistent DAG extension of G: can we derive a score-based check if $i \to j$ is linear or nonlinear in this extension based on G?

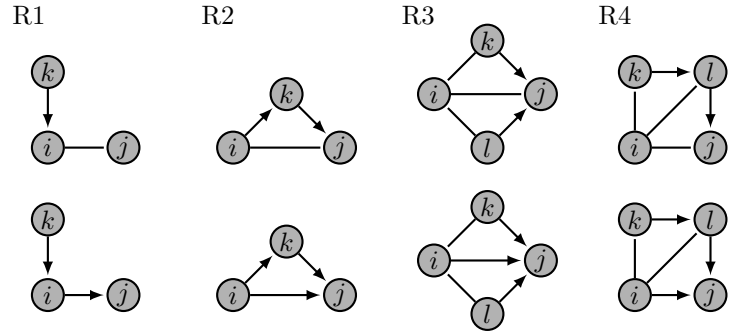


Figure 3.6: Orientation rules R1-R4 for Markov equivalence classes with background knowledge (Meek, 1995). If there is an edge constellation as in the top row, $i - j$ is oriented as $i \to j$ when closing orientations under R1-R4.

Interestingly, the answer to both questions is yes (cf. Lemma 5) and can be derived from a related theory on how background knowledge on specific edge orientations restricts the Markov equivalence class. It was shown in Theorems 2 and 4 in Meek (1995) that for a pattern P of a DAG, consistent background knowledge $\mathcal{K}$ (in our case: additional knowledge on edge orientations due to nonlinear functions in the PLSEM) can be incorporated by simply orienting these edges in P and closing orientations under a set of four sound and complete graphical orientation rules R1-R4, which are depicted in Figure 3.6. The resulting PDAG, which we denote by $G_{P,\mathcal{K}}$, is maximally oriented with respect to the set of all Markov equivalent DAGs with edge orientations that comply with the background knowledge.

It is important to note that we generally do not obtain $G_{\mathscr{D}(\mathbb{P})}$ if we simply add all nonlinear edges in D^0 as background knowledge $\mathcal{K}$ and close orientations under R1-R4. The resulting maximally oriented PDAG $G_{P,\mathcal{K}}$ is typically not equal to $G_{\mathscr{D}(\mathbb{P})}$. For an example, consider D_1 in Figure 3.1 and denote by P_1 its pattern. For $\mathcal{K} = \{1 \to 2\}$ we obtain the PDAG $G_{P_1,\mathcal{K}}$ with undirected edge $1 - 3$. But $1 \to 3$ in $G_{\mathscr{D}(\mathbb{P})}$ by Definition 1 as $\mathscr{D}(\mathbb{P}) = \{D_1, D_2\}$. This illustrates that we have to add all edges to $\mathcal{K}$ that are nonlinear in a DAG in $\mathscr{D}(\mathbb{P})$ in which they are covered ($1 \to 3$ is covered and nonlinear in D_2).

Lemma 5. *Let P be the pattern of a DAG and $\mathcal{K}$ a consistent set of background knowledge (not containing directed edges of P). Let $G_{P,\mathcal{K}}$ denote the maximally oriented graph with respect to P and $\mathcal{K}$ with orientations closed under R1-R4.*

(a) Edge $i \to j$ in $\mathcal{K}$ is not covered in any of the consistent DAG extensions of $G_{P,\mathcal{K}}$ if and only if $G_{P,\mathcal{K}} = G_{P,\mathcal{K}\setminus\{i\to j\}}$.

(b) If $G_{P,\mathcal{K}} \neq G_{P,\mathcal{K}\setminus\{i\to j\}}$, there exists a consistent DAG extension of $G_{P,\mathcal{K}}$ in which $\mathrm{pa}_{G_{P,\mathcal{K}}}(j) \setminus \{i\}$ is a cover for $i \to j$.

A proof is given in Appendix 3.A.6. By construction, $G_{P,\mathcal{K}} = G_{P,\mathcal{K}\setminus\{i\to j\}}$ if and only if the orientation of $i \to j$ in $G_{P,\mathcal{K}\setminus\{i\to j\}}$ is implied by one of R1-R4 applied to $G_{P,\mathcal{K}}$ with undirected edge $i - j$. Hence, Lemma 5 (a) answers (Q1) as it provides a simple graphical criterion to check whether $i \to j$ in $G_{P,\mathcal{K}}$ is covered in one of the consistent DAG extensions of $G_{P,\mathcal{K}}$ based on $G_{P,\mathcal{K}}$ only. Note that part (a) is closely related to Section 5 in Andersson et al. (1997), where the authors construct the CPDAG (representing the Markov equivalence class) from a given DAG by removing edge orientations that are not implied by a set of graphical orientation rules, which contain R1-R3 in Figure 3.6. Lemma 5 (b) answers (Q2): it allows us to implement a score-based check whether $i \to j$ is linear or nonlinear in a DAG extension of $G_{P,\mathcal{K}}$ in which it is covered by simply reading off the parents of j in $G_{P,\mathcal{K}}$ and use them as a cover for $i \to j$. Details are given in Remark 5.

We now propose the following iterative estimation procedure for $G_{\mathscr{D}(\mathbb{P})}$: let $D^0 \in \mathscr{D}(\mathbb{P})$ be given, P denote its pattern and define $\mathcal{K}_1 := \mathcal{K}_1^{\text{init}} \cup \mathcal{K}_1^{\text{nonl}}$, where $\mathcal{K}_1^{\text{init}}$ contains all directed edges in D^0 that are undirected in P and $\mathcal{K}_1^{\text{nonl}} := \emptyset$. By construction, $G_{P,\mathcal{K}_1} = D^0$. For $k \geq 1$, in each iteration k to $k+1$, we apply Lemma 5 (a) and use R1-R4 to select an edge $\{i \to j\} \in \mathcal{K}_k^{\text{init}}$ ($i \to j$ in $G_{P,\mathcal{K}_k}$) that is covered in a consistent DAG extension of $G_{P,\mathcal{K}_k}$ (that is, not implied by any of R1-R4). If $\mathcal{K}_k^{\text{init}} = \emptyset$ or no such edge exists, we stop and output $G_{P,\mathcal{K}_k}$. Else, we check whether $i \to j$ is linear or nonlinear in a consistent DAG extension in which it is covered and construct a new set of background knowledge $\mathcal{K}_{k+1} := \mathcal{K}_{k+1}^{\text{init}} \cup \mathcal{K}_{k+1}^{\text{nonl}} \subseteq \mathcal{K}_k$ according to the following rules:

Case 1: If $i \to j$ is linear, $\mathcal{K}_{k+1}^{\text{nonl}} = \mathcal{K}_k^{\text{nonl}}$ and $\mathcal{K}_{k+1}^{\text{init}} = \mathcal{K}_k^{\text{init}} \setminus \{i \to j\}$.
Case 2: If $i \to j$ is nonlinear, $\mathcal{K}_{k+1}^{\text{nonl}} = \mathcal{K}_k^{\text{nonl}} \cup \{i \to j\}$ and $\mathcal{K}_{k+1}^{\text{init}} = \mathcal{K}_k^{\text{init}} \setminus \{i \to j\}$.

In particular, by construction, Case 1 implies that $i - j$ in all $G_{P,\mathcal{K}_l}$ for $l > k$, whereas Case 2 fixes the orientation $i \to j$ in all $G_{P,\mathcal{K}_l}$ for $l > k$.

Lemma 6. *Let $\{\mathcal{K}_k\}_k$ be constructed as described above. Then, the sequence of maximally oriented PDAGs $\{G_{P,\mathcal{K}_k}\}_k$ converges to $G_{\mathscr{D}(\mathbb{P})}$.*

The result is proven in Appendix 3.A.6 and illustrated in Figure 3.7. As in both cases, $|\mathcal{K}_{k+1}^{\text{init}}| = |\mathcal{K}_k^{\text{init}}| - 1$, $\{G_{P,\mathcal{K}_k}\}_k$ converges to $G_{\mathscr{D}(\mathbb{P})}$ after at most $|\mathcal{K}_1^{\text{init}}|$ iterations, where $|\mathcal{K}_1^{\text{init}}|$ is the number of undirected edges in P.

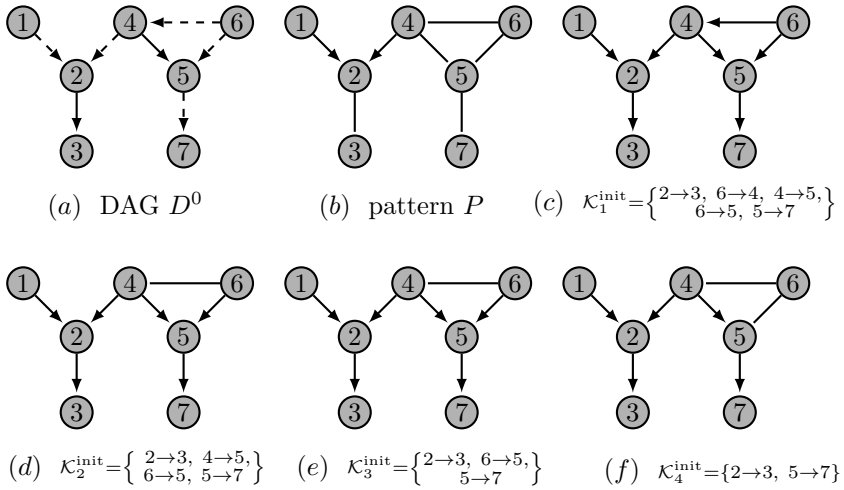


Figure 3.7: Illustration of Algorithm 3. (a) DAG D^0 with linear edges (dashed) and nonlinear edges (solid). (b) step 2: pattern P of D^0. (c) step 3: directed edges in D^0 that are undirected in P are added to $\mathcal{K}_1^{\text{init}}$. By construction, $\widehat{G}_{n,p}$ is equal to D^0. (c)-(f) steps 4-12: $4 \leftarrow 6$ is covered and linear in (c), hence, orientation is removed in $\widehat{G}_{n,p}$ in (d). $4 \to 5$ is covered and nonlinear in a consistent DAG extension of (d), hence, orientation is fixed in $\widehat{G}_{n,p}$ in (e). $6 \to 5$ is covered and linear in a consistent DAG extension of (e), hence, orientation is removed in $\widehat{G}_{n,p}$ in (f). As both edges in $\mathcal{K}_4^{\text{init}}$ are implied by R1 in (f), they are not covered in any of the consistent DAG extensions of $\widehat{G}_{n,p}$ in (f). Concludingly, $\widehat{G}_{n,p} = G_{\mathscr{D}(\mathbb{P})}$ in (f).

Remark 5. *Let $\{i \to j\} \in \mathcal{K}_k^{\text{init}}$ be the edge chosen in iteration k to $k+1$. By Lemma 5 (b), $S := \text{pa}_{G_{P,\mathcal{K}_k}}(j) \setminus \{i\}$ is a cover of $i \to j$ in one of the consistent DAG extensions of $G_{P,\mathcal{K}_k}$. From that, we easily obtain a score-based version: we simply regress X_i on X_S and X_j on $X_{S\cup\{i\}}$ to obtain the estimates $\hat{\sigma}_i, \hat{\sigma}_j$ of the standard deviations of the residuals at nodes i and j for $i \to j$. Similarly, we regress X_i on $X_{S\cup\{j\}}$ and X_j on X_S to obtain the estimates $\hat{\sigma}'_i, \hat{\sigma}'_j$ for $i \leftarrow j$. If the estimated score difference $|\log(\hat{\sigma}'_i) + \log(\hat{\sigma}'_j) - \log(\hat{\sigma}_i) - \log(\hat{\sigma}_j)|$ is smaller than α, we conclude that $i \to j$ is linear, else, nonlinear. The pseudo-code of the score-based procedure is provided in Algorithm 3. It outputs an estimate $\widehat{G}_{n,p}$ of $G_{\mathscr{D}(\mathbb{P})}$ based on n samples $X^{(1)}, ..., X^{(n)}$ and $D^0 \in \mathscr{D}(\mathbb{P})$.*

Algorithm 3 computeGDPX

Initialize $\widehat{G}_{n,p} \leftarrow D^0$, $k \leftarrow 1$, $\mathcal{K}_1^{\text{init}} \leftarrow \emptyset$ and $\mathcal{K}_1^{\text{nonl}} \leftarrow \emptyset$.
Construct the pattern P of D^0.
Add directed edges in D^0 that are undirected in P to $\mathcal{K}_1^{\text{init}}$.
while There is $i \to j$ in $\mathcal{K}_k^{\text{init}}$, such that its orientation is not implied by rules R1, R2, R3 or R4 applied to $\widehat{G}_{n,p}$ with undirected edge $i - j$ **do**
  Use $\text{pa}_{\widehat{G}_{n,p}}(j) \setminus \{i\}$ to cover $i \to j$ and estimate the standard deviations $\hat{\sigma}_i, \hat{\sigma}_j, \hat{\sigma}'_i, \hat{\sigma}'_j$ of the residuals as described in Remark 5.
  if $|\log(\hat{\sigma}'_i) + \log(\hat{\sigma}'_j) - \log(\hat{\sigma}_i) - \log(\hat{\sigma}_j)| < \alpha$ **then**
    Set $\mathcal{K}_{k+1}^{\text{init}} \leftarrow \mathcal{K}_k^{\text{init}} \setminus \{i \to j\}$ and replace $i \to j$ by $i - j$ in $\widehat{G}_{n,p}$.
  else
    Set $\mathcal{K}_{k+1}^{\text{init}} \leftarrow \mathcal{K}_k^{\text{init}} \setminus \{i \to j\}$ and keep $i \to j$ in $\widehat{G}_{n,p}$.
  end if
  $k \leftarrow k + 1$.
end while
return Estimated PDAG $\widehat{G}_{n,p}$ representing $\mathscr{D}(\mathbb{P})$.

A major advantage of Algorithm 3 is that it can be implemented based on one adjacency matrix only that is updated in every iteration.

Theorem 9. *Under Assumption 1 and $\xi_p \geq \xi_0 > 0$, for any $\alpha \in (0, \xi_0)$,*

$$\mathbb{P}\left[\widehat{G}_{n,p} = G_{\mathscr{D}(\mathbb{P})}\right] \to 1 \qquad (n \to \infty).$$

Proof. The correctness of Algorithm 3 is proved in Lemma 6. The consistency of the score-based estimation follows from the proof of Theorem 8. □

3.4 Simulations

In this section we empirically analyze the performance of `computeGDPX` (Algorithm 3) in various settings. Consider $\mathbb{P}$ that has been generated by a faithful PLSEM with known DAG D^0. The goal is to estimate the corresponding distribution equivalence class $\mathscr{D}(\mathbb{P})$ based on D^0 and samples of $\mathbb{P}$. In Section 3.4.1, we describe the simulation setting. We then briefly comment on a population version of Algorithm 3 in Section 3.4.2, which is used to obtain the underlying true distribution equivalence class $\mathscr{D}(\mathbb{P})$. In

the subsequent sections we examine the role of the tuning parameter α (Section 3.4.3), the performance in low- and high-dimensional settings (Section 3.4.4) and the computation time (Section 3.4.5).

3.4.1 Simulation setting and implementation details

Throughout the section, let p denote the number of variables, n the number of samples, n_{rep} the number of repetitions of an experiment, p_{c} the probability to connect two nodes by an edge and p_{lin} the probability that an edge is linear. For each experiment we generate n_{rep} random true DAGs D^0 with the function `randomDAG` in the R-package `pcalg` (Kalisch et al., 2012) with parameters `n =` p and `prob =` p_{c}. For each of the random DAGs, we generate n samples of $\mathbb{P}$ from a PLSEM with edge functions chosen as follows: with probability p_{lin}, $f_{j,i}(x) = \alpha_{j,i} \cdot x$ is linear with $\alpha_{j,i}$ randomly drawn from $[-1.5, -0.5] \cup [0.5, 1.5]$. Otherwise, $f_{j,i}(x)$ is nonlinear and randomly drawn from the set $\{c_0 \cdot \cos(c_1 \cdot (x - c_2)), c_0 \cdot \tanh(c_1 \cdot (x - c_2))\}$ to have a mix of monotone and non-monotone functions in the PLSEM. In order to be able to empirically support our theoretical findings we choose the parameters $c_0 \sim \text{Unif}([-2, -1] \cup [1, 2]), c_1 \sim \text{Unif}([1, 2])$ and $c_2 \sim \text{Unif}([-\pi/3, \pi/3])$ such that the nonlinear functions are "sufficiently nonlinear" and not too close to linear functions. Exemplary randomly generated nonlinear functions are shown in Figure 3.8. The noise variables satisfy $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_j^2 \sim \text{Unif}([1, 2])$ for source nodes (nodes with empty parental set) and $\sigma_j^2 \sim \text{Unif}([1/4, 1/2])$ otherwise.

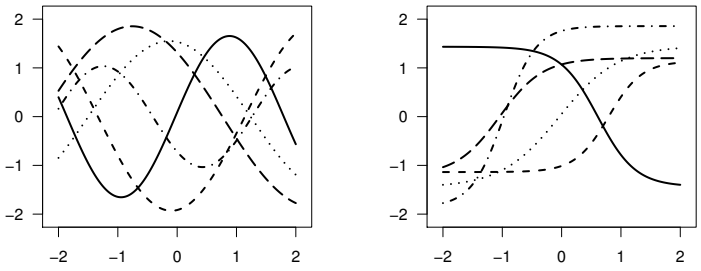


Figure 3.8: Exemplary nonlinear functions used in simulated PLSEMs

In order to estimate the residuals in step 5 of `computeGDPX`, we use additive model fitting based on the `R`-package `mgcv` with default settings (Wood,

2003, 2006). The basis dimension for each smooth term is set to 6.

There exists no state-of-the-art method that we can compare our algorithm with. In principle, given D^0, we can estimate the corresponding PLSEMs for all DAGs in the Markov equivalence class of D^0 and compute their scores. This also gives us an estimate for $\mathscr{D}(\mathbb{P})$, but as explained in Section 3.3.2, is less efficient than `computeGDPX`. We therefore only evaluate how accurately `computeGDPX` estimates $G_{\mathscr{D}(\mathbb{P})}$. For that, let $G_{\mathscr{D}(\mathbb{P})}$ and $\hat{G}$ denote the true and estimated graphical representations of $\mathscr{D}(\mathbb{P})$, respectively. We count

(i) the number of edges that are undirected in $G_{\mathscr{D}(\mathbb{P})}$ but directed in $\hat{G}$ ("falsely kept orientations")

(ii) the number of edges that are directed in $G_{\mathscr{D}(\mathbb{P})}$ but undirected in $\hat{G}$ ("falsely removed orientations").

Note that as we assume faithfulness, all DAGs in $\mathscr{D}(\mathbb{P})$ have the same CPDAG. By construction, `computeGDPX` does not falsely remove orientations on the directed part of the CPDAG as all these edges are not covered in any of the consistent DAG extensions. To obtain the percentages shown in Figures 3.9 to 3.11 we therefore only divide by the number of undirected edges in the CPDAG. The percentages then reflect a measure for the fraction of "correct score-based decisions".

3.4.2 Reference method for the true distribution equivalence class $\mathscr{D}(\mathbb{P})$

To be able to characterize the true $\mathscr{D}(\mathbb{P})$ based on D^0 and the corresponding PLSEM we assume that for each $i \in \{1, ..., p\}$, the functions in the set $\{\partial_i^2 f_{j,i} : j \text{ is a child of } i \text{ in } D^0 \text{ and } f_{j,i} \text{ is nonlinear}\}_i$ are linearly independent for the PLSEM with DAG D^0 that generates $\mathbb{P}$. As all functions in our simulations are randomly drawn (cf. Section 3.4.1), the assumption is satisfied with probability one for D^0 and the corresponding edge functions.

This additional assumption rules out cases where nonlinear effects in D^0 exactly cancel out over different paths and hence excludes cases as in Figure 3.4 where nonlinear edges may be reversed. In particular, it allows us to use Theorem 11 to obtain $G_{\mathscr{D}(\mathbb{P})}$ only based on D^0 and knowledge of the functions in the corresponding PLSEM: first, we use Theorem 11 (c) to construct the set $\mathcal{V}$. For all nodes i in D^0, corresponding sets of nonlinear

children C_i (as defined in Appendix 3.A.5) and $k \neq i$, we add (i, k) to $\mathcal{V}$ if k is a descendant of a node in C_i. In principle, we now apply Algorithm 3, but instead of the score-based decision in steps 6-9, we use the set $\mathcal{V}$ to decide about edge orientations. Let $i \to j$ be the edge chosen in step 4 and D one of the consistent DAG extensions in which $i \to j$ is covered. If $(i, j) \in \mathcal{V}$, by Theorem 11 (d) and Remark 10, $i \to j$ in all DAGs of a PLSEM that generates $\mathbb{P}$. Hence, in particular, $i \to j$ in all DAGs in $\mathscr{D}(\mathbb{P})$ and by definition, $i \to j$ in $G_{\mathscr{D}(\mathbb{P})}$. If $(i, j) \notin \mathcal{V}$, by Lemma 7, the DAG D' that differs from D only by reversing $i \to j$ is in $\mathscr{D}(\mathbb{P})$. Hence, by definition, $i - j$ in $G_{\mathscr{D}(\mathbb{P})}$.

3.4.3 The role of α for varying sample size

In `computeGDPX`, the score-based decision if a selected covered edge is linear or nonlinear is based on a comparison of the absolute difference of the expected negative log-likelihood scores of two models with a parameter α. In Figure 3.9, we empirically analyze the dependence of $\hat{G}$ on α for sparse graphs and different sample sizes.

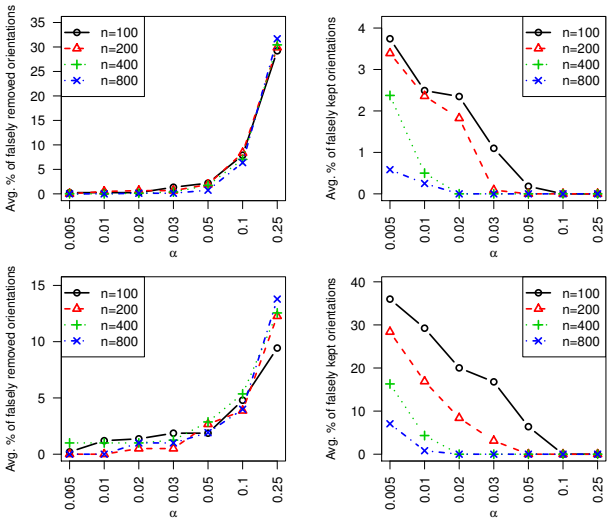


Figure 3.9: Performance of `computeGDPX` for varying sample sizes and values of α (x-axis) in sparse DAGs with $p_{\text{lin}} = 0.2$ (top) and $p_{\text{lin}} = 0.8$ (bottom). Parameters: $p = 10$, $n_{\text{rep}} = 100$ and $p_{\text{c}} = 2/9$ (expected number of edges: 10).

Optimally, one would choose α close to ξ_p, see equation (3.10), but ξ_p depends on the setting (number of variables, sparsity of the DAG, degree of nonlinearity of the nonlinear functions, etc.) and is unknown. In practice, the parameter α reflects a measure of how conservative the estimate $\hat{G}$ of $G_{\mathscr{D}(\mathbb{P})}$ is (in the sense of how many causal statements can be made). For example, choosing α large results in a conservative estimate $\hat{G}$ with many undirected edges (a large set $\mathscr{D}(\mathbb{P})$ of equivalent DAGs). `computeGDPX` exhibits a good performance for a wide range of values of α. In particular, as the sample size increases, choosing α small results in very accurate estimates $\hat{G}$ of $G_{\mathscr{D}(\mathbb{P})}$. The sparsity of the DAG does not strongly influence the results, see Figure 3.10.

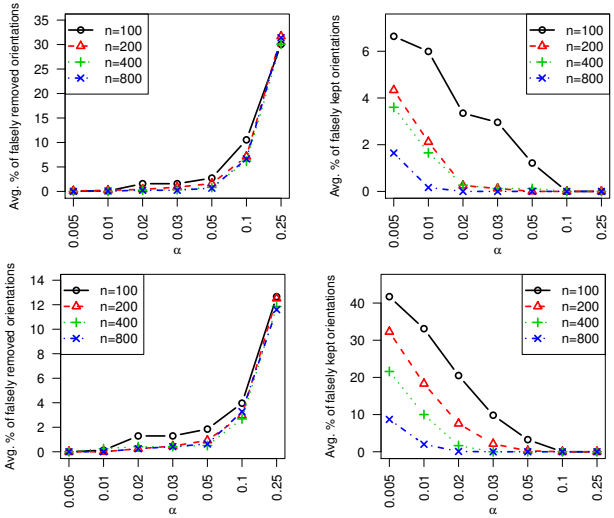


Figure 3.10: Performance of `computeGDPX` for varying sample sizes and values of α (x-axis) in dense DAGs for $p_{\text{lin}} = 0.2$ (top) and $p_{\text{lin}} = 0.8$ (bottom). Parameters: $p = 10$, $n_{\text{rep}} = 100$ and $p_{\text{c}} = 6/9$ (expected number of edges: 30).

3.4.4 The dependence on p: low- and high-dimensional setting

From the fact that `computeGDPX` only relies on local score computations, we expect that its performance does not strongly depend on the number of variables p as long as the neighborhood sizes in the DAGs (the node de-

grees) are similar for different values of p. We simulate $n_{\text{rep}} = 100$ random DAGs with $p = 10$, $p = 100$ and $p = 1000$ nodes, respectively. Moreover, we set $p_{\text{c}} = 2/(p-1)$ which results in an expected number of p edges and an expected node degree of 2 for all settings. As demonstrated in Figure 3.11, the accuracy of `computeGDPX` with respect to varying values of α is barely affected by the number of variables p. In particular, `computeGDPX` exhibits a good performance even in high-dimensional settings with $p = 1000$ and sample sizes in the hundreds. The same conclusions hold for $p_{\text{c}} = 6/(p-1)$ with an expected node degree of 6 (not shown).

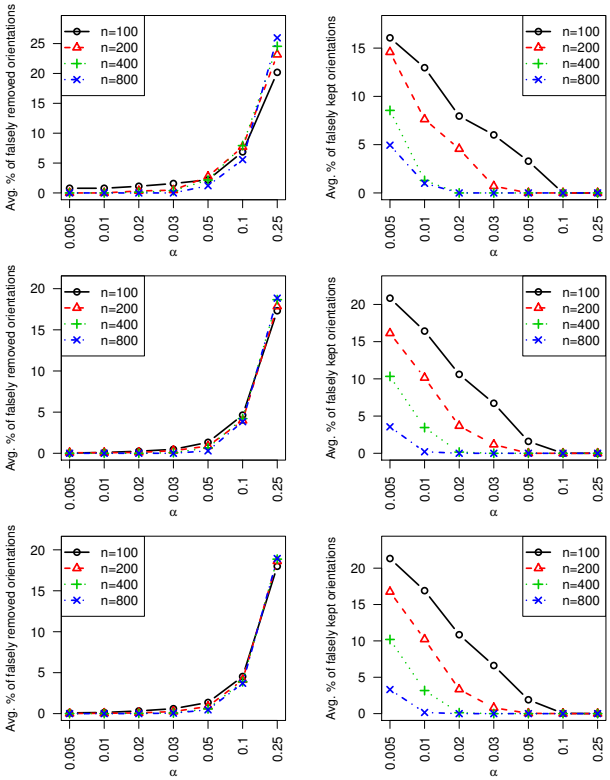


Figure 3.11: Performance of `computeGDPX` for varying sample sizes and values of α (x-axis) for $p = 10$ (top), $p = 100$ (middle) and $p = 1000$ (bottom). Parameters: $p_{\text{lin}} = 0.5$, $n_{\text{rep}} = 100$ and $p_{\text{c}} = 2/(p-1)$ (expected number of edges: p).

3.4.5 Computation time

Lastly, we analyze the computation time of `computeGDPX` depending on the number of variables p and sparsity p_c. We examine the following two scenarios: (i) most of the functions in the PLSEM are nonlinear ($p_{\text{lin}} = 0.2$) and (ii) the worst-case scenario (w.r.t. computation time) where all the functions in the PLSEM are linear ($p_{\text{lin}} = 1$) and $\mathscr{D}(\mathbb{P})$ is equal to the Markov equivalence class ($G_{\mathscr{D}(\mathbb{P})}$ is equal to the CPDAG).

For all combinations of $p \in \{10, 20, 50, 100, 250, 500, 1000, 2000, 5000\}$ and $p_c \in \{2/(p-1), 8/(p-1)\}$ and scenarios (i) and (ii), we measure the time consumption of `computeGDPX` for $n = 400$ and $\alpha = 0.05$. In the scenario where all the functions are linear, we additionally compare it to `dag2cpdag` in the R-package `pcalg`, which constructs the CPDAG based on iterative application of R1-R3 in Figure 3.6. The median CPU times are shown in Table 3.1.

Table 3.1: Median CPU times [s] for `computeGDPX` and for `dag2cpdag` that iteratively applies R1 to R3 in Figure 3.6. $n_{\text{rep}} = 100$ repetitions for $p_{\text{lin}} = 0.2$ and $n_{\text{rep}} = 20$ repetitions for $p_{\text{lin}} = 1$.

	$p_{\text{lin}} = 0.2$		$p_{\text{lin}} = 1$			
	`computeGDPX`		`computeGDPX`		`dag2cpdag`	
$\mathbb{E}[\lvert\text{edges}\rvert]$	p	$4p$	p	$4p$	p	$4p$
$p = 10$	0.092	0.785	0.157	1.101	0.007	0.005
$p = 20$	0.150	0.105	0.174	0.162	0.006	0.006
$p = 50$	0.300	0.164	0.332	0.223	0.008	0.009
$p = 100$	0.604	0.281	0.665	0.325	0.014	0.016
$p = 250$	1.446	0.630	1.740	0.717	0.072	0.087
$p = 500$	2.705	1.253	3.486	1.523	0.395	0.599
$p = 1000$	5.616	2.513	6.603	2.974	3.464	4.231
$p = 2000$	11.504	5.380	13.493	6.331	25.463	31.591
$p = 5000$	29.226	16.276	35.094	18.462	400.324	591.574

`computeGDPX` is able to estimate $G_{\mathscr{D}(\mathbb{P})}$ in less than a minute even if the number of variables is in the thousands. In general, the speed of our implementation heavily depends on the sparsity of the DAGs. This can be seen from the case with $p = 10$ and expected number of edges 40. In this setting the DAGs are almost fully connected. This in turn implies that not many of the edges are fixed due to v-structures and a lot of score-based tests have to be performed. On the other hand, if the underlying DAGs are

sparse, we observe that `computeGDPX` even outperforms `dag2cpdag` with respect to computation time if the number of variables is large. Note that this only holds for sparse DAGs. In general, `dag2cpdag` is much faster than our implementation (not shown).

3.5 Conclusions

We comprehensively characterized the identifiability of partially linear structural equation models with Gaussian noise (PLSEMs) from various perspectives. First, we proved that under faithfulness we obtain graphical and transformational characterizations of distribution equivalent DAGs similar to well-known characterizations of Markov equivalence classes of DAGs. More generally, we demonstrated that reinterpreting PLSEMs as PLSEM-functions leads to an interesting geometric characterization of all PLSEMs that generate the same distribution $\mathbb{P}$, as they can all be expressed as constant rotations of each other. Therefrom we derived a precise condition how PLSEM-functions (and hence also how single nonlinear additive components in PLSEMs) restrict the set of potential causal orderings of the variables and showed how it can be leveraged to conclude about the causal relations of specific pairs of variables under mild additional assumptions. The theoretical results were complemented with an efficient algorithm that finds all equivalent DAGs to a given DAG or PLSEM. We proved its high-dimensional consistency and evaluated its performance on simulated data.

These characterizations of PLSEMs (and corresponding DAGs) that generate the same distribution $\mathbb{P}$ are crucial for further algorithmic developments in structure learning, for example in the spirit of Castelo and Kocka (2003), or for Monte Carlo sampling in Bayesian settings, see a related discussion in Andersson et al. (1997, Section 1).

Appendix 3.A Technical results and proofs

This appendix contains detailed specifications and proofs of the theorems in Chapter 3. The order of the presentation matches the one in the previous sections. Figure 3.12 gives an overview of the dependency structure of the different theorems.

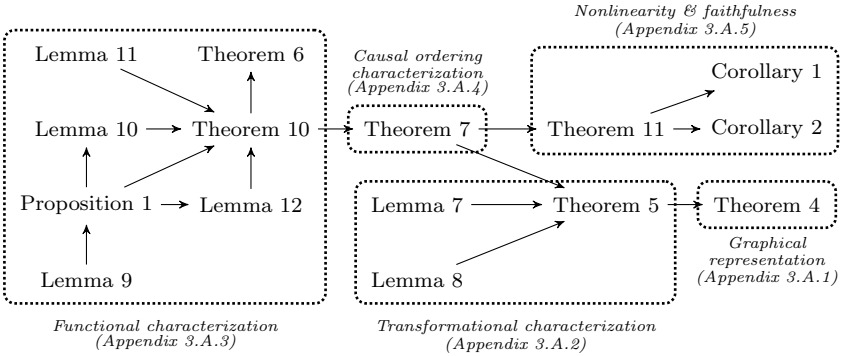


Figure 3.12: Proof structure for the characterization results in Section 3.2. The proofs for Section 3.3 are given in Appendix 3.A.6 (not depicted).

3.A.1 Proof of the graphical characterization (Theorem 4)

Proof. By definition, $\mathscr{D}(\mathbb{P})$ is a subset of the set of all consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$. It remains to show, that the set of all consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$ is a subset of $\mathscr{D}(\mathbb{P})$. Suppose there is a consistent DAG extension $\tilde{D}$ of $G_{\mathscr{D}(\mathbb{P})}$ such that $\tilde{D} \notin \mathscr{D}(\mathbb{P})$. Let $D \in \mathscr{D}(\mathbb{P})$. As both, D and $\tilde{D}$ are consistent DAG extensions of $G_{\mathscr{D}(\mathbb{P})}$, they have the same skeleton and v-structures and are Markov equivalent. Hence, there exists a sequence of distinct covered edge reversals transforming D into $\tilde{D}$ (Chickering, 1995, Theorem 2). Let us denote the sequence of traversed DAGs by $D = D_1, ..., D_m = \tilde{D}$. If all covered edge reversals are linear, $\tilde{D} \in \mathscr{D}(\mathbb{P})$ by Theorem 5 (a), which contradicts the assumption. Therefore, there is at least one covered nonlinear edge reversal in this sequence. Without loss of generality, for $1 \leq r \leq m-1$, let the edge reversal of $i \to j$ to $i \leftarrow j$ between D_r and D_{r+1} be the first covered nonlinear edge reversal in the

above sequence. First note that as the sequence of covered edge reversals is distinct, $i \to j$ in D and $i \leftarrow j$ in $\tilde{D}$. Moreover, as D_r is obtained from D by a sequence of covered linear edge reversals, $D_r \in \mathscr{D}(\mathbb{P})$ by Theorem 5 (a). Again, by Theorem 5 (a), as $D_r \in \mathscr{D}(\mathbb{P})$ and $i \to j$ is covered and nonlinear in D_r, $i \to j$ for all DAGs $D' \in \mathscr{D}(\mathbb{P})$. Therefore, by Definition 1, $i \to j$ in $G_{\mathscr{D}(\mathbb{P})}$ which contradicts the assumption that $\tilde{D}$ is a consistent DAG extension of $G_{\mathscr{D}(\mathbb{P})}$. □

3.A.2 Proof of the transformational characterization (Theorem 5)

Proof. (a): By Lemma 8 there exists a unique PLSEM with DAG D that generates $\mathbb{P}$. Let F denote the function that corresponds to this PLSEM as defined in Section 3.2.2. Without loss of generality let us assume that $\mathrm{D}F$ is lower triangular. Furthermore, as $i \to j$ is covered in D, no other child of i is an ancestor of j and we can assume that $j = i + 1$. The differential $\mathrm{D}F$ is of the form

$$\begin{pmatrix} \mathrm{Var}(\varepsilon_1)^{-1/2} & 0 & \dots & \dots & \dots & 0 \\ \partial_1 F_2 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathrm{Var}(\varepsilon_i)^{-1/2} & 0 & \ddots & \vdots \\ \vdots & \ddots & \partial_i F_{i+1} & \mathrm{Var}(\varepsilon_{i+1})^{-1/2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \partial_1 F_p & \dots & \dots & \dots & \dots & \mathrm{Var}(\varepsilon_p)^{-1/2} \end{pmatrix}.$$

Let us write $v = (\mathrm{D}F)^{-1}\partial_i^2 F$, that is, $\partial_i^2 F = \mathrm{D}F v$. As $\mathrm{D}F$ is lower triangular with $\left(\mathrm{Var}(\varepsilon_i)^{-1/2}\right)_{i=1,\dots,p}$ on the diagonal we get $v_1, \dots, v_i = 0$ and $v_{i+1} = \mathrm{Var}(\varepsilon_{i+1})^{1/2}\partial_i^2 F_{i+1}$. Hence,

$$e_{i+1}^t (\mathrm{D}F)^{-1}\partial_i^2 F = \mathrm{Var}(\varepsilon_{i+1})^{1/2}\partial_i^2 F_{i+1}.$$

Now recall that by definition of F,

$$\partial_i^2 F_{i+1} = -\frac{1}{\mathrm{Var}(\varepsilon_{i+1})^{1/2}}\partial_i^2 f_{i+1,i}(x_i).$$

By combining these two equations,

$$e_{i+1}^t (\mathrm{D}F)^{-1}\partial_i^2 F = -\partial_i^2 f_{i+1,i}(x_i). \tag{3.11}$$

By Lemma 7, the edge can be reversed if and only if $(i, i+1) \notin \mathcal{V}$, which by definition of $\mathcal{V}$ is the case if and only if

$$e_{i+1}^t (\mathrm{D}F)^{-1} \partial_i^2 F \equiv 0.$$

By equation (3.11) this is the case if and only if $\partial_i^2 f_{i+1,i}(x_i) \equiv 0$. Hence, the edge can be reversed if and only if the edge is linear. This concludes the proof of the "if and only if" statement.

If the edge $i \to i+1$ is nonlinear, we can argue analogously as above that $(i, i+1) \in \mathcal{V}$. By Theorem 7, all causal orderings of PLSEMs that generate $\mathbb{P}$ satisfy $\sigma(i) < \sigma(i+1)$. As, by definition, $\mathbb{P}$ is faithful to all DAGs in $\mathscr{D}(\mathbb{P})$, they all have the same skeleton. Hence, $i \to i+1$ in all DAGs in $\mathscr{D}(\mathbb{P})$.

(b): As $D, D' \in \mathscr{D}(\mathbb{P})$, D' is Markov equivalent to D. Hence, there exists a sequence of distinct covered edge reversals transforming D into D' (Chickering, 1995, Theorem 2). Let us denote the sequence of traversed DAGs by $D = D_1, \ldots, D_m = D'$. By part (a), we are done if we can show that each DAG D_r in this sequence lies in $\mathscr{D}(\mathbb{P})$. We prove this by induction. So let us assume $D_r \in \mathscr{D}(\mathbb{P})$ with $r < m$. Then D_{r+1} only differs from D_r by the reversal of a covered edge, w.l.o.g. $i \to j$ in D_r and $j \to i$ in D_{r+1}. By construction, all covered edge reversals are distinct, hence, $j \to i$ in D'. Define the set $\mathcal{V}$ as in Theorem 7. As $D, D' \in \mathscr{D}(\mathbb{P})$, by Theorem 7, $(i, j) \notin \mathcal{V}$. Hence by Lemma 7 we immediately get that $D_{r+1} \in \mathscr{D}(\mathbb{P})$. Moreover, by Theorem 5 (a), $i \to j$ is linear. This concludes the proof. □

Lemma 7. *Let $D \in \mathscr{D}(\mathbb{P})$. Let $i \to j$ be a covered edge in D. Let D' be a DAG that differs from D only by reversing $i \to j$. Let F be a PLSEM-function of $\mathbb{P}$ and define $\mathcal{V}$ as in equation (3.8). Then $D' \in \mathscr{D}(\mathbb{P})$ if and only if $(i, j) \notin \mathcal{V}$.*

Proof. "⇒": Let $D' \in \mathscr{D}(\mathbb{P})$ and $(i, j) \in \mathcal{V}$. Consider a causal ordering σ of D'. As $j \to i$ in D', $\sigma(j) < \sigma(i)$. By Theorem 7 this leads to a contradiction. Hence if $D' \in \mathscr{D}(\mathbb{P})$, then $(i, j) \notin \mathcal{V}$.

"⇐": Let $(i, j) \notin \mathcal{V}$. Let σ be a causal ordering of D. As $i \to j$ is covered in D, no other child of i is an ancestor of j in D. Hence without loss of generality we can assume that $\sigma(j) = \sigma(i) + 1$. Define σ' as the

permutation with i and j switched, that is,

$$\sigma'(k) = \begin{cases} \sigma(k) & k \notin \{i, j\} \\ \sigma(j) & k = i \\ \sigma(i) & k = j. \end{cases}$$

Note that as $\sigma(j) = \sigma(i)+1$, the causal orderings of other pairs of variables are unaffected. Hence,

$$\sigma(k) < \sigma(l) \iff \sigma'(k) < \sigma'(l) \text{ for all } k, l \text{ with } \{k, l\} \neq \{i, j\}. \tag{3.12}$$

As σ is a causal ordering of a PLSEM that generates $\mathbb{P}$, by Theorem 7,

$$\sigma(k) < \sigma(l) \text{ for all } (k, l) \in \mathcal{V}. \tag{3.13}$$

We want to show that the same holds for σ'. Let $(k, l) \in \mathcal{V}$. As $(i, j) \notin \mathcal{V}$, $(k, l) \neq (i, j)$. Hence, by equations (3.12) and (3.13), $\sigma'(k) < \sigma'(l)$. This proves that

$$\sigma'(k) < \sigma'(l) \text{ for all } (k, l) \in \mathcal{V}.$$

By Theorem 7, σ' is a causal ordering of a PLSEM that generates $\mathbb{P}$. Consider the DAG $\tilde{D}$ of this PLSEM. Then $\mathbb{P}$ is Markov with respect to $\tilde{D}$ and by Proposition 17 of Peters et al. (2014), $\mathbb{P}$ satisfies causal minimality with respect to $\tilde{D}$. By Lemma 1 in Chickering (1995), $\mathbb{P}$ is Markov and faithful with respect to D' and we know that σ' is a causal ordering of both $\tilde{D}$ and D'. Now we want to show that this implies $\tilde{D} = D'$. Without loss of generality assume $\sigma' = \mathrm{Id}$. First, we want to show that $\mathrm{pa}_{\tilde{D}}(l) \supseteq \mathrm{pa}_{D'}(l)$ for all l. Fix l. Consider the parental set $\mathrm{pa}_{\tilde{D}}(l)$ of l in $\tilde{D}$ and let k be a parent of l in D' but not in $\tilde{D}$. As $\sigma' = \mathrm{Id}$ is a causal ordering of D', $k < l$, and as $\sigma' = \mathrm{Id}$ is a causal ordering of $\tilde{D}$ as well, k is not a descendant of l in $\tilde{D}$. As $\mathbb{P}$ is Markov with respect to $\tilde{D}$,

$$X_l \perp\!\!\!\perp X_k | X_{\mathrm{pa}_{\tilde{D}}(l)}.$$

Hence, as $\mathbb{P}$ is faithful to D', l and k are d-separated by $\mathrm{pa}_{\tilde{D}}(l)$ in D'. But k is a parent of l in D', contradiction. Hence $\mathrm{pa}_{\tilde{D}}(l) \supseteq \mathrm{pa}_{D'}(l)$ for all l. $\mathbb{P}$ satisfies causal minimality with respect to $\tilde{D}$, hence $\mathrm{pa}_{\tilde{D}}(l) = \mathrm{pa}_{D'}(l)$ for all l. This proves $\tilde{D} = D'$. Therefore, there exists a PLSEM with DAG D' that generates $\mathbb{P}$ and $\mathbb{P}$ is faithful with respect to D'. By definition, $D' \in \mathscr{D}(\mathbb{P})$. □

Lemma 8. *Let $\mathbb{P}$ be generated by a PLSEM. Let $D \in \mathscr{D}(\mathbb{P})$. Then there exists a unique PLSEM (unique set of intercepts, edge functions and Gaussian error variances) with DAG D that generates $\mathbb{P}$.*

Proof. By definition of the distribution equivalence class $\mathscr{D}(\mathbb{P})$ there exists such a PLSEM with DAG D that generates $\mathbb{P}$. Now we will show that this PLSEM is unique. Consider another PLSEM with DAG D that generates $\mathbb{P}$. For a given node j we have

$$\mu_j + \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) = \mathbb{E}[X_j | X_{\mathrm{pa}_D(j)}] = \tilde{\mu}_j + \sum_{i \in \mathrm{pa}_D(j)} \tilde{f}_{j,i}(X_i).$$

By definition of PLSEMs, the expectations of the $f_{j,i}(X_i)$ and $\tilde{f}_{j,i}(X_i)$ are zero, hence we have $\mu_j = \tilde{\mu}_j$. As $\sigma_k > 0$ for all $k \in \{1, \ldots, p\}$, the density of X is positive on $\mathbb{R}^p$. Recall that by definition, $f_{j,i}$ and $\tilde{f}_{j,i}$ are continuous. Hence, for all $x \in \mathbb{R}^p$,

$$\begin{aligned}\sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(x_i) &= \lim_{\delta \to 0} E\left[\sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) | X \in B_\delta(x)\right] \\ &= \lim_{\delta \to 0} E\left[\sum_{i \in \mathrm{pa}_D(j)} \tilde{f}_{j,i}(X_i) | X \in B_\delta(x)\right] = \sum_{i \in \mathrm{pa}_D(j)} \tilde{f}_{j,i}(x_i),\end{aligned}$$

where $B_\delta(x)$ denotes the closed ball around x with radius δ. Take an arbitrary $i \in \mathrm{pa}_D(j)$. By taking the derivative with respect to x_i on both sides of the equation we obtain

$$f'_{j,i}(x_i) = \tilde{f}'_{j,i}(x_i).$$

Hence there exists a constant c such that

$$f_{j,i}(x_i) = c + \tilde{f}_{j,i}(x_i).$$

By definition of PLSEMs, we have $\mathbb{E}[f_{j,i}(X_i)] = 0$ and $\mathbb{E}[\tilde{f}_{j,i}(X_i)] = 0$. Hence, $c = 0$ and $f_{j,i} = \tilde{f}_{j,i}$ for all $i \in \mathrm{pa}_D(j)$. We just showed that $\mu_j = \tilde{\mu}_j$ and $f_{j,i} = \tilde{f}_{j,i}$. It remains to show that $\sigma_j = \tilde{\sigma}_j$:

$$\begin{aligned}\sigma_j^2 &= \mathrm{Var}\left(X_j - \mu_j - \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i)\right) \\ &= \mathrm{Var}\left(X_j - \tilde{\mu}_j - \sum_{i \in \mathrm{pa}_D(j)} \tilde{f}_{j,i}(X_i)\right) = \tilde{\sigma}_j^2.\end{aligned}$$

Hence, the intercepts, edge functions and Gaussian error variances of both PLSEMs are equal, which concludes the proof. □

3.A.3 Proof of the functional characterization (Theorem 6)

In the following, let $\mathbb{P}$ be generated by a PLSEM.

Definition 2 (PLSEM-functions)**.** *We call $F : \mathbb{R}^p \to \mathbb{R}^p$ a PLSEM-function of $\mathbb{P}$ if there exists a PLSEM that generates $\mathbb{P}$ such that F can be written as in equation* (3.3)*.*

Remark 6. *For a PLSEM-function F of $\mathbb{P}$ we can retrieve the unique corresponding PLSEM (i.e. the unique DAG, unique set of intercepts, edge functions and Gaussian error variances) through equations* (3.4)*,* (3.5) *and* (3.6)*.*

Proposition 1. *A function $F : \mathbb{R}^p \to \mathbb{R}^p$ is a PLSEM-function of $\mathbb{P}$ if and only if*

1. *F is twice continuously differentiable,*
2. *$\partial_k \partial_l F \equiv 0$ for all $k \neq l$,*
3. *there exists a permutation σ such that $(\mathrm{D}F_{i\sigma^{-1}(j)})_{ij}$ is lower triangular with constant positive entries on the diagonal.*
4. *If $X \sim \mathbb{P}$, then $F(X) \sim \mathcal{N}(0, Id_p)$,*

We call a permutation σ that satisfies (3) a causal ordering of the PLSEM-function F. Define the directed graph D, the functions $f_{j,i}$, $\sigma_j^2 = \mathrm{Var}(\varepsilon_j)$ and μ_j through equations (3.4) – (3.6)*. The first condition reflects that the functions $f_{j,i}$ are twice continuously differentiable. The second condition reflects that the functions $f_{j,i}$ depend on x_i only. The third condition ensures that the directed graph D is acyclic and that the variances of all ε_j are strictly positive. The last condition ensures that the distribution generated by this PLSEM is $\mathbb{P}$.*

Proof. "⇒" By definition of a PLSEM and equation (3.3).
"⇐": Without loss of generality let us assume that the indices are ordered such that $\sigma = \mathrm{Id}$. By (3), $\mathrm{D}F$ is lower triangular with constant positive

entries on the diagonal. Let $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$ and define $X := F^{-1}(Z)$. Using (4), we obtain $X \sim \mathbb{P}$. Use (1) and (2) and Lemma 9 for each component of F_j, i.e. decompose $F_j(x) = \tilde{\mu}_j + \sum_i \tilde{f}_{j,i}(x_i)$ with twice continuously differentiable functions $\tilde{f}_{j,i}$. Here we choose $\tilde{\mu}_j$ and the $\tilde{f}_{j,i}$ (i.e. the constants) such that $\mathbb{E}[\tilde{f}_{j,i}(X_i)] \equiv 0$ for all $j \neq i$ and $\tilde{f}_{j,j}$ such that $\tilde{f}_{j,j}(0) = 0$. We define the parental sets $\mathrm{pa}(j) := \{i \neq j : \tilde{f}_{j,i} \not\equiv 0\}$. As $\mathrm{D}F$ is lower triangular, $\mathrm{pa}(j) \subseteq \{1, \dots, j-1\}$, hence the directed graph D defined by these parental sets is acyclic. As $\mathrm{D}F$ has constant positive entries on the diagonal, $\partial_j F_j$ is constant, and we can define the error variances $\sigma_j^2 := 1/\left(\partial_j F_j\right)^2 > 0$. Furthermore, we define the functions $f_{j,i}(x_i) := -\sigma_j \tilde{f}_{j,i}(x_i)$ that only depend on x_i and constants $\mu_j := -\sigma_j \tilde{\mu}_j$. To sum it up, we have the following relations:

$$F_j(x) = \frac{1}{\sigma_j}\left(x_j - \mu_j - \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(x_i)\right),$$

with DAG D, $f_{j,i} \not\equiv 0$, $\mathbb{E}[f_{j,i}(X_i)] = 0$ for all $i \in \mathrm{pa}_D(j)$. Using that $F(X) = Z \sim \mathcal{N}(0, \mathrm{Id}_p)$,

$$X_j = \mu_j + \sum_{i \in \mathrm{pa}_D(j)} f_{j,i}(X_i) + \sigma_j Z_j.$$

By defining the Gaussian errors $\varepsilon_j := \sigma_j Z_j$, it is immediate to see that $\sigma_j, f_{j,i}, D$ define a PLSEM that generates $\mathbb{P}$. □

Theorem 10 (Functional characterization)**.** *Let F be a PLSEM-function of $\mathbb{P}$. Let σ be a permutation. Define Π^{σ}_{i+1} as the linear projection on the space $\langle \partial_{\sigma^{-1}(i+1)} F, \dots, \partial_{\sigma^{-1}(p)} F \rangle$ and $\Pi^{\sigma}_{p+1} := 0 \in \mathbb{R}^{p \times p}$. Let $G : \mathbb{R}^p \to \mathbb{R}^p$. Then G is a PLSEM-function of $\mathbb{P}$ with causal ordering σ if and only if*

$$\left(Id - \Pi^{\sigma}_{i+1}\right) \partial^2_{\sigma^{-1}(i)} F \equiv 0, \qquad i = 1, \dots, p,$$

and

$$G_i = \left(\frac{(Id - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)} F}{\|(Id - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)} F\|_2}\right)^t F. \tag{3.14}$$

In that case, the matrices Π^{σ}_{i+1} and the vectors $(Id - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)} F$ are constant.

Remark 7. *This theorem tells us that every potential causal ordering satisfies $\left(Id - \Pi^{\sigma}_{i+1}\right) \partial^2_{\sigma^{-1}(i)} F \equiv 0$, $i = 1, \dots, p$, and contains a concrete formula to compute the unique PLSEM-function for this given causal ordering. Furthermore, every causal ordering that satisfies that condition gives*

rise to a corresponding PLSEM-function by equation (3.14). *As F is a PLSEM-function, its Jacobian* DF *is invertible. This then implies that* $\|(Id - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F\|_2 > 0$, *hence, equation* (3.14) *is well-defined. Given the PLSEM-function, we can retrieve the unique corresponding PLSEM (i.e. the unique DAG, unique set of intercepts, edge functions and Gaussian error variances) through equations* (3.4) – (3.6).

Remark 8. *If G is a PLSEM-function, from this theorem it follows that the vectors*

$$\left(\frac{(Id - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F}{\|(Id - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F\|_2}\right)^t \qquad i = 1, \ldots, p,$$

are constant in x, have unit norm and are orthogonal for $i = 1, \ldots, p$. Hence the row-wise concatenation of these vectors for $i = 1, \ldots, p$ forms an orthogonal matrix O and by equation (3.14)*: $G = OF$.*

Proof. "⇒". Let G be a PLSEM-function of $\mathbb{P}$ with causal ordering σ. Without loss of generality let $\sigma = \mathrm{Id}$, i.e. without loss of generality we assume that DG is lower triangular. We write Π_{i+1} instead of Π^{σ}_{i+1} for brevity. Define $J := G(F^{-1})$. By Proposition 1 (3), $\det \mathrm{D}G$ and $\det \mathrm{D}F$ are constant. Hence, $\det \mathrm{D}J$ is constant, too. Furthermore, by Proposition 1 (4), $J(\varepsilon) \sim \mathcal{N}(0, \mathrm{Id}_p)$ for $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_p)$. By Lemma 11 we obtain

$$\|G(x)\|_2^2 = \|F(x)\|_2^2 \text{ for all } x \in \mathbb{R}^p.$$

By differentiating on both sides,

$$G^t \mathrm{D}G = F^t \mathrm{D}F.$$

We assumed without loss of generality that $\sigma = \mathrm{Id}$, hence by Proposition 1 the differential DG is lower triangular and the diagonal entries $c_i := \partial_i G_i$ are positive. Hence we can recursively solve for $i = 1, \ldots, p$ and obtain

$$G_i = \frac{1}{c_i}\left(F^t \partial_i F - \sum_{j>i} G_j \partial_i G_j\right) \tag{3.15}$$

Using induction, we will show that

$$c_i G_i = F^t(\mathrm{Id} - \Pi_{i+1})\partial_i F, \tag{3.16}$$

that $c_i = \|(\mathrm{Id} - \Pi_{i+1})\partial_i F\|_2$, that the matrix Π_{i+1} is constant and that the vectors $(\mathrm{Id} - \Pi_{i+1})\partial_i^2 F \equiv 0$ for $i = 1, \ldots, p$. By using equation (3.15) we

immediately obtain equation (3.16) for $i = p$ and by Lemma 10 we obtain $(\mathrm{Id} - \Pi_{p+1})\partial_p^2 F = \partial_p^2 F \equiv 0$. Hence,

$$c_p^2 = c_p \partial_p G_p = (\partial_p F)^t \partial_p F = \|(\mathrm{Id} - \Pi_{p+1})\partial_p F\|_2^2.$$

Furthermore, $(\mathrm{Id} - \Pi_{p+1})\partial_p F$ is a constant vector and hence by definition the matrix Π_p is constant. Now let us assume $c_j G_j = F^t(\mathrm{Id} - \Pi_{j+1})\partial_j F$, $c_j = \|(\mathrm{Id} - \Pi_{j+1})\partial_j F\|_2$, that the matrix Π_j is constant and that the vectors $(\mathrm{Id} - \Pi_{j+1})\partial_j^2 F \equiv 0$ for all $p \geq j > i \geq 1$. We want to prove these statements for $j = i$. By using equation (3.15) and the induction assumption we can rewrite $c_i G_i$,

$$\begin{aligned}
c_i G_i &= F^t \partial_i F - \sum_{j>i} G_j \partial_i G_j \\
&= F^t \partial_i F - \sum_{j>i} \frac{F^t(\mathrm{Id} - \Pi_{j+1})\partial_j F}{c_j} \frac{\partial_i F^t(\mathrm{Id} - \Pi_{j+1})\partial_j F}{c_j} \\
&= F^t \left(\mathrm{Id} - \sum_{j>i} \frac{(\mathrm{Id} - \Pi_{j+1})\partial_j F}{\|(\mathrm{Id} - \Pi_{j+1})\partial_j F\|_2} \frac{((\mathrm{Id} - \Pi_{j+1})\partial_j F)^t}{\|(\mathrm{Id} - \Pi_{j+1})\partial_j F\|_2} \right) \partial_i F \\
&= F^t \left(\mathrm{Id} - \Pi_{i+1}\right) \partial_i F.
\end{aligned}$$

By Lemma 10 we get $\left(\mathrm{Id} - \Pi_{i+1}\right) \partial_i^2 F \equiv 0$ and hence

$$\begin{aligned}
c_i^2 &= c_i \partial_i G_i \\
&= \partial_i \left(F^t \left(\mathrm{Id} - \Pi_{i+1}\right) \partial_i F \right) \\
&= \partial_i F^t \left(\mathrm{Id} - \Pi_{i+1}\right) \partial_i F + 0 \\
&= \|(\mathrm{Id} - \Pi_{i+1})\partial_i F\|_2^2.
\end{aligned}$$

It remains to show that Π_i is constant. We already proved that the vector $\left(\mathrm{Id} - \Pi_{i+1}\right) \partial_i^2 F \equiv 0$. Π_{i+1} is constant by induction assumption. Thus, $\partial_i \left(\left(\mathrm{Id} - \Pi_{i+1}\right) \partial_i F\right) \equiv 0$. By Proposition 1 (2), $\partial_i F$ depends only on x_i. Hence the vector $\left(\mathrm{Id} - \Pi_{j+1}\right) \partial_j F$ is constant for $j = i$. By induction assumption we also know that this is true for all $j > i$. By definition, we know that

$$\Pi_i = \sum_{j \geq i} \frac{(\mathrm{Id} - \Pi_{j+1})\partial_j F}{\|(\mathrm{Id} - \Pi_{j+1})\partial_j F\|_2} \frac{((\mathrm{Id} - \Pi_{j+1})\partial_j F)^t}{\|(\mathrm{Id} - \Pi_{j+1})\partial_j F\|_2}.$$

As shown, the quantities on the right-hand side are constant. This concludes the proof by induction.

"$\Leftarrow$" We will show 1) $-$ 4) of Proposition 1 to prove that G is a PLSEM-function of $\mathbb{P}$. By Lemma 12, the vectors $(\mathrm{Id}-\Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F$ are constant. As F is twice continuously differentiable, G is twice differentiable as well. This proves 1). By part 2) of Proposition 1, $\partial_k\partial_l F = 0$ for all $k \neq l$. Recall that the vector $(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F$ is constant. Let $k \neq l$. Hence $\partial_k\partial_l G_i = (\partial_k\partial_l F)^t(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F/\|(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F\|_2 = 0$. This proves that for all $k \neq l$, $\partial_k\partial_l G = 0$, i.e., part 2) of Proposition 1. Now we want to show that $(\mathrm{D}G_{i\sigma^{-1}(j)})_{ij}$ is lower triangular. By construction, $\mathrm{D}G_{i\sigma^{-1}(j)} = \partial_{\sigma^{-1}(j)}G_i = \partial_{\sigma^{-1}(j)}F^t(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F = 0$ for all $j > i$ as by definition $\partial_{\sigma^{-1}(j)}F^t(\mathrm{Id} - \Pi^{\sigma}_{i+1}) = 0$. It remains to be shown that $(\mathrm{D}G_{i\sigma^{-1}(j)})_{ij}$ has positive constant entries on the diagonal. Recall that by assumption $(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial^2_{\sigma^{-1}(i)}F \equiv 0$ for $i = 1, \ldots, p$ and that the vector $(\mathrm{Id}-\Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F$ is constant. The vector is non-zero as $\mathrm{D}F$ is invertible. Hence,

$$\mathrm{D}G_{i\sigma^{-1}(i)} = \frac{\partial_{\sigma^{-1}(i)}F^t(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F}{\|(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F\|_2} = \frac{\|(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F\|_2^2}{\|(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F\|_2}.$$

Thus $\mathrm{D}G_{i\sigma^{-1}(i)}$ is constant. This proves 3). Let $X \sim \mathbb{P}$. Now it remains to be shown that $G(X) \sim \mathcal{N}(0, \mathrm{Id})$. To this end, note that by definition of Π^{σ}_{j+1}, the vectors

$$(\mathrm{Id} - \Pi^{\sigma}_{j+1})\partial_{\sigma^{-1}(j)}F, \qquad j = 1, \ldots, p,$$

are orthogonal. As shown above, these vectors are constant and non-zero, therefore $(\mathrm{Id} - \Pi^{\sigma}_{j+1})\partial_{\sigma^{-1}(j)}F, j = 1, \ldots, p$, is an orthogonal basis of $\mathbb{R}^p$. Therefore, $\|F(x)\|_2^2 = \|G(x)\|_2^2$ for all $x \in \mathbb{R}^p$. As $\det \mathrm{D}G$ is constant, we have by the change of variables formula that $|\det \mathrm{D}G| = |\det \mathrm{D}F|$ (probability densities integrate to one). Hence, again by the change of variables formula, $G(X) \sim \mathcal{N}(0, \mathrm{Id})$, which is 4). This concludes the proof of the "if and only if" statement.

Lemma 12 proves that in that case, the matrices Π^{σ}_{i+1} and the vectors $(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)}F$ are constant. This concludes the proof. □

Lemma 9. *Let $F : \mathbb{R}^p \mapsto \mathbb{R}$ be twice continuously differentiable. If $\partial_k\partial_l F \equiv 0$ for all $l \neq k$, then F can be written in the form*

$$F(x) = c + g_1(x_1) + \ldots + g_p(x_p). \tag{3.17}$$

In this case, the functions $g_i(x_i)$ are unique up to constants and twice continuously differentiable.

Proof. Fix an arbitrary $y \in \mathbb{R}^p$. We use Taylor:

$$\begin{aligned} F(x) - F(y) &= \int_{y_1}^{x_1} \partial_1 F(z, x_2, ..., x_p)\mathrm{d}z + ... \\ &\qquad + \int_{y_p}^{x_p} \partial_p F(y_1, y_2, ..., y_{p-1}, z)\mathrm{d}z \\ &= \int_{y_1}^{x_1} \partial_1 F(z, 0, ..., 0)\mathrm{d}z + ... + \int_{y_p}^{x_p} \partial_p F(0, 0, ..., 0, z)\mathrm{d}z \end{aligned}$$

In the second line we used that $\partial_k \partial_l F \equiv 0$ for all $l \neq k$. Now we can define

$$g_i(x_i) = \int_{y_i}^{x_i} \partial_i F(0, ..., 0, z, 0, ..., 0)\mathrm{d}z,$$

which proves equation (3.17) with constant $c = F(y)$. Furthermore, as

$$\partial_i F = \partial_i g_i(x_i),$$

the g_i are unique up to constants. This completes the proof. □

Lemma 10. *Let $F, G : \mathbb{R}^p \to \mathbb{R}^p$ be PLSEM-functions of $\mathbb{P}$. Let Π be a constant projection matrix. If $c_i G_i = F^t(Id - \Pi)\partial_i F$ for a constant c_i, then*

$$(Id - \Pi)\partial_i^2 F \equiv 0.$$

Proof. Recall Proposition 1. We will use properties 1), 2) and 3) of F and G in the proof. As G is a PLSEM-function, by Proposition 1

$$\partial_j F^t(\mathrm{Id} - \Pi)\partial_i^2 F = c_i \partial_i \partial_j G_i = 0 \text{ for all } j \neq i.$$

As Π is a projection matrix, $(\mathrm{Id}-\Pi)^t(\mathrm{Id}-\Pi) = (\mathrm{Id}-\Pi)(\mathrm{Id}-\Pi) = (\mathrm{Id}-\Pi)$, which implies

$$((\mathrm{Id} - \Pi)\partial_j F)^t (\mathrm{Id} - \Pi)\partial_i^2 F = 0 \text{ for all } j \neq i. \tag{3.18}$$

As F is a PLSEM-function, by Proposition 1, there exists a permutation σ such that $(\mathrm{D}F_{k\sigma^{-1}(l)})_{kl}$ is lower triangular with constant positive entries on the diagonal. Without loss of generality let us assume that $\sigma = \mathrm{Id}$, i.e, that the variables x_i are ordered such that $\mathrm{D}F$ is lower triangular with constant positive entries on the diagonal. Hence $\langle \partial_p F, \ldots, \partial_{i+1} F\rangle = \langle e_p, \ldots, e_{i+1}\rangle$, and $\partial_i^2 F \in \langle e_p, \ldots, e_{i+1}\rangle$. Furthermore, $\partial_i^2 F \in \langle \partial_p F, \ldots, \partial_{i+1} F\rangle$ which implies $(\mathrm{Id}-\Pi)\partial_i^2 F \in \langle (\mathrm{Id}-\Pi)\partial_p F, \ldots, (\mathrm{Id}-\Pi)\partial_{i+1} F\rangle$. By equation (3.18), $(\mathrm{Id} - \Pi)\partial_i^2 F \equiv 0$. □

Lemma 11. *Let $\varepsilon \sim \mathcal{N}(0, Id_p)$ and $J : \mathbb{R}^p \to \mathbb{R}^p$ be a $\mathcal{C}^1$-diffeomorphism such that $\det \mathrm{D}J$ is constant and $J(\varepsilon) \sim \mathcal{N}(0, Id_p)$. Then, $|\det \mathrm{D}J| = 1$ and*

$$\|J(e)\|_2 = \|e\|_2 \text{ for all } e \in \mathbb{R}^p.$$

Proof. By assumption, for any Borel set $A \subset \mathbb{R}^p$, $\mathbb{P}[\varepsilon \in A] = \mathbb{P}[J^{-1}(\varepsilon) \in A]$. Hence, by the change of variables formula,

$$\begin{aligned}\int_A \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{\|\varepsilon\|_2^2}{2}\right) \mathrm{d}\varepsilon &= \int_{J(A)} \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{\|\varepsilon\|_2^2}{2}\right) \mathrm{d}\varepsilon \\ &= \int_A |\det \mathrm{D}J| \frac{1}{\left(\sqrt{2\pi}\right)^p} \exp\left(-\frac{\|J(\varepsilon)\|_2^2}{2}\right) \mathrm{d}\varepsilon.\end{aligned}$$

By continuity of J, for all $\varepsilon \in \mathbb{R}^p$, $\|\varepsilon\|_2^2 = \|J(\varepsilon)\|_2^2 - 2\log(|\det \mathrm{D}J|)$. As J is a diffeomorphism there exists an ε^0 such that $J(\varepsilon^0) = 0$. This immediately implies $\log(|\det \mathrm{D}J|) \leq 0$. Analogously, for $\varepsilon = 0$ we obtain $\log(|\det \mathrm{D}J|) \geq 0$. Hence, $\log(|\det \mathrm{D}J|) = 0$ and for all $\varepsilon \in \mathbb{R}^p$, it holds that $\|\varepsilon\|_2^2 = \|J(\varepsilon)\|_2^2$. This concludes the proof. □

Lemma 12. *Let F be a PLSEM-function and σ be a permutation on $\{1, \ldots, p\}$. Let*

$$\left(Id - \Pi_{i+1}^{\sigma}\right) \partial^2_{\sigma^{-1}(i)} F \equiv 0, \text{ for } i = 1, \ldots, p, \tag{3.19}$$

where Π_{i+1}^{σ} denotes the linear projection on $\langle \partial_{\sigma^{-1}(i+1)} F, \ldots, \partial_{\sigma^{-1}(p)} F \rangle$ and $\Pi_{p+1}^{\sigma} = 0 \in \mathbb{R}^{p \times p}$. Then the matrices Π_{i+1}^{σ} and vectors $(Id - \Pi_{i+1}^{\sigma}) \partial_{\sigma^{-1}(i)} F$ are constant for $i = 1, \ldots, p$.

Proof. Let us first show that the projection matrices Π_{i+1}^{σ} are constant. For $i = p$ the claim is trivial as $\Pi_{p+1}^{\sigma} \equiv 0$ and hence by equation (3.19), $\partial^2_{\sigma^{-1}(p)} F \equiv 0$. For arbitrary i, equation (3.19) implies that

$$\partial^2_{\sigma^{-1}(i)} F \in \langle \partial_{\sigma^{-1}(i+1)} F, \ldots, \partial_{\sigma^{-1}(p)} F \rangle.$$

Furthermore, as F is a PLSEM-function, by Proposition 1, $\partial_{\sigma^{-1}(i)} F$ only depends on $x_{\sigma^{-1}(i)}$. Using these two facts it now follows inductively that the linear spaces $\langle \partial_{\sigma^{-1}(i+1)} F, \ldots, \partial_{\sigma^{-1}(p)} F \rangle$, $i = p-1, \ldots, 1$ are constant in x. Hence the linear projections on these spaces are constant matrices. This proves that the matrices Π_{i+1}^{σ}, $i = 1, \ldots, p$ are constant. Now we want to show that the vectors $(\mathrm{Id} - \Pi_{i+1}^{\sigma}) \partial_{\sigma^{-1}(i)} F$,

$i = 1, \dots, p$ are constant. Recall that $\partial_{\sigma^{-1}(i)} F$ only depends on $x_{\sigma^{-1}(i)}$, so $(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)} F$ only depends on $x_{\sigma^{-1}(i)}$. It therefore suffices to show that $\partial_{\sigma^{-1}(i)}(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)} F = 0$. By equation (3.19) and as the matrix Π^{σ}_{i+1} is constant, $\partial_{\sigma^{-1}(i)}(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial_{\sigma^{-1}(i)} F = (\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial^2_{\sigma^{-1}(i)} F = 0$. This concludes the proof. □

3.A.4 Proof of characterization via causal orderings (Theorem 7)

Remark 9. *Slight abuse of notation. A priori, $\mathcal{V}$ might depend on the concrete choice of F. However by Theorem 7, the set of permutations $\{\sigma : \sigma(i) < \sigma(k)$ for all $(i,k) \in \mathcal{V}\}$ does only depend on $\mathbb{P}$.*

Proof. Let Π^{σ}_{i+1} be the linear projection on $\langle \partial_{\sigma^{-1}(i+1)} F, \dots, \partial_{\sigma^{-1}(p)} F \rangle$ and $\Pi^{\sigma}_{p+1} := 0 \in \mathbb{R}^{p \times p}$. By some algebra,

$$
\begin{aligned}
&(\mathrm{Id} - \Pi^{\sigma}_{i+1})\partial^2_{\sigma^{-1}(i)} F \equiv 0 \text{ for all } i \\
\iff &\partial^2_{\sigma^{-1}(i)} F(x) \in \langle \partial_{\sigma^{-1}(i+1)} F(x), \dots, \partial_{\sigma^{-1}(p)} F(x) \rangle \text{ for all } i, \forall x \in \mathbb{R}^p \\
\iff &e^t_{\sigma^{-1}(j)} (\mathrm{D}F)^{-1} \partial^2_{\sigma^{-1}(i)} F \equiv 0 \text{ for all } j \leq i \\
\iff &j > i \text{ for all } e^t_{\sigma^{-1}(j)} (\mathrm{D}F)^{-1} \partial^2_{\sigma^{-1}(i)} F \not\equiv 0 \\
\iff &\sigma(j) > \sigma(i) \text{ for all } e^t_j (\mathrm{D}F)^{-1} \partial^2_i F \not\equiv 0.
\end{aligned}
$$

Here, $e_j, j = 1, \dots, p$, denotes the standard basis of $\mathbb{R}^p$. By invoking Theorem 10 and Remark 7 the assertion follows. □

3.A.5 Interplay of nonlinearity and faithfulness (Corollaries 1 & 2)

Let $C_i := \{j \text{ child of } i \text{ in } D : \partial^2_i f_{j,i} \not\equiv 0\}$ be the set of nonlinear children of i and consider the following condition for a fix $i \in \{1, \dots, p\}$:

(C1) The functions $f_{j,i}, j \in C_i$ are linearly independent.

Theorem 11 (Identifiability of nonlinear descendants under minor assumptions)**.** *Consider a PLSEM with DAG D that generates $\mathbb{P}$ and the corresponding PLSEM-function F. Define $\mathcal{V}$ as in equation* (3.8) *and C_i as above. Fix $i \in \{1, \dots, p\}$. Then:*

(a) *Assume (C1) and let $j \in C_i$. Then $(i,j) \in \mathcal{V}$.*

(b) *Let $(i,j) \in \mathcal{V}$. Then j is a descendant of i in every DAG D' of a PLSEM that generates $\mathbb{P}$.*

(c) *Let $\mathbb{P}$ be faithful to D and assume (C1). Consider a descendant k of i, $k \neq i$. Then $(i,k) \in \mathcal{V}$ if and only if k is descendant of one of the nonlinear children in C_i.*

(d) *Define $\tilde{\mathcal{V}}$ as the transitive closure of $\mathcal{V}$. Let $k \neq i$. Then $(i,k) \in \tilde{\mathcal{V}}$ if and only if k is a descendant of i in every DAG D' of a PLSEM that generates $\mathbb{P}$.*

Remark 10. *We follow the convention that i is a descendant of itself. If (C1) holds for all i, then from (c) it follows that $\tilde{\mathcal{V}} = \mathcal{V}$. In particular, by (d), k is descendant of one of the nonlinear children in C_i if and only if k is a descendant of i in every DAG D' of a PLSEM that generates $\mathbb{P}$.*

Proof. We will first show (a) and (c). Recall Proposition 1. Without loss of generality let us assume that $\sigma = \mathrm{Id}$, i.e. that $\mathrm{D}F$ is lower triangular with constant positive entries on the diagonal. Furthermore, $\partial_i^2 F_l = 0$ for all $l \leq i$. Using this we obtain $(\mathrm{D}F)^{-1}\partial_i^2 F = (\mathrm{D}F)^{-1}_{\bullet,(i+1):p}\partial_i^2 F_{(i+1):p}$, and

$$\begin{aligned} & e_k^t(\mathrm{D}F)^{-1}\partial_i^2 F && \equiv 0 \qquad (3.20)\\ \iff & e_k^t(\mathrm{D}F)^{-1}_{\bullet,(i+1):p}\partial_i^2 F_{(i+1):p} && \equiv 0. \end{aligned}$$

Here the subindex $\bullet$ denotes all rows $1:p$. In the next step, we want to prove that

$$\langle \partial_i^2 F_{(i+1):p}(x_i)\rangle_{x_i\in\mathbb{R}} = \langle (e_j)_{(i+1):p}\rangle_{j\in C_i}. \qquad (3.21)$$

Note that as the components $1,\ldots,i$ of $\partial_i^2 F$ and $e_j, j \in C_i$ are zero, this is equivalent to showing that

$$\langle \partial_i^2 F(x_i)\rangle_{x_i\in\mathbb{R}} = \langle e_j\rangle_{j\in C_i}.$$

As $\partial_i^2 F_l \equiv 0$ for all $l \notin C_i$, we have

$$\langle \partial_i^2 F(x_i)\rangle_{x_i\in\mathbb{R}} \subseteq \langle e_j\rangle_{j\in C_i}.$$

Let $\gamma \in \langle e_j\rangle_{j\in C_i}$, $\gamma \neq 0$.

$$\left(\partial_i^2 F(x_i)\right)^t \gamma = \sum_{j\in C_i} \partial_i^2 F_j(x_i)\gamma_j$$

As the nonlinear children in C_i are linearly independent, there exists an $x_i \in \mathbb{R}$ such that $\sum_j \partial_i^2 F_j(x_i)\gamma_j \neq 0$. Hence there exists no non-zero vector γ in $\langle e_j \rangle_{j \in C_i}$ that is orthogonal to $\langle \partial_i^2 F(x_i) \rangle_{x_i \in \mathbb{R}}$ and hence

$$\langle \partial_i^2 F(x_i) \rangle_{x_i \in \mathbb{R}} = \langle e_j \rangle_{j \in C_i}.$$

As discussed above, this proves equation (3.21). By Proposition 1, each column l of $\mathrm{D}F$ is a function of x_l (i.e. constant in $x_1, \ldots, x_{l-1}, x_{l+1}, \ldots x_p$). Hence, $(\mathrm{D}F)^{-1}_{\bullet,(i+1):p}$ is a function of $x_{i+1}, \ldots, x_p$. In particular, it is constant in x_i. Now we can continue:

$$\begin{aligned}
& e_k^t (\mathrm{D}F)^{-1}_{\bullet,(i+1):p} \partial_i^2 F_{i+1:p} \equiv 0 \qquad (3.22)\\
\iff & e_k^t (\mathrm{D}F)^{-1}_{\bullet,(i+1):p}(x_{i+1}, \ldots, x_p) \partial_i^2 F_{i+1:p}(x_i) = 0 \text{ for all } x \in \mathbb{R}^p \\
\iff & e_k^t (\mathrm{D}F)^{-1}_{\bullet,(i+1):p}(x_{i+1}, \ldots, x_p)(e_j)_{(i+1):p} = 0 \text{ for all } j \in C_i, x \in \mathbb{R}^p \\
\iff & e_k^t (\mathrm{D}F)^{-1}(x) e_j = 0 \text{ for all } j \in C_i, x \in \mathbb{R}^p
\end{aligned}$$

As $\mathrm{D}F$ is lower triangular with positive entries on the diagonal, $(\mathrm{D}F)^{-1}$ is lower triangular, too, with non-zero entries on the diagonal. Hence, $e_k^t (\mathrm{D}F)^{-1}(x) e_k \not\equiv 0$. So if $k \in C_i$, by equation (3.22),

$$e_k^t (\mathrm{D}F)^{-1}_{\bullet,(i+1):p} \partial_i^2 F_{i+1:p} \not\equiv 0.$$

By equation (3.20), $e_k^t (\mathrm{D}F)^{-1} \partial_i^2 F \not\equiv 0$ and hence by definition of $\mathcal{V}$, $(i,k) \in \mathcal{V}$. This proves (a).

Let $Z \sim \mathcal{N}(0, \mathrm{Id}_p)$. Let $X = F^{-1}(Z)$. By Proposition 1, $X \sim \mathbb{P}$. Note that $X_k = e_k^t F^{-1}(Z)$. We denote the partial derivative with respect to z_j by ∂_j^z. Note that x_k is constant in z_j if and only if $\partial_j^z e_k^t F^{-1}(z) = e_k^t (\mathrm{D}F)^{-1}(F^{-1}(z)) e_j \equiv 0$. Fix j. As there are bijective relationships between x and z and between $x_{1:(j-1)}$ and $z_{1:(j-1)}$,

$$\begin{aligned}
& e_k^t (\mathrm{D}F)^{-1}(x) e_j = 0 \quad \text{for all } x \in \mathbb{R}^p \qquad (3.23)\\
\iff & e_k^t (\mathrm{D}F)^{-1}(F^{-1}(z)) e_j = 0 \quad \text{for all } z \in \mathbb{R}^p \\
\iff & x_k = e_k^t F^{-1}(z) \text{ is constant in } z_j \\
\iff & x_k = e_k^t F^{-1}(z) \text{ is a function of } z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_k \\
& \Rightarrow X_k \perp\!\!\!\perp Z_j | Z_1, \ldots, Z_{j-1} \\
& \Rightarrow X_k \perp\!\!\!\perp Z_j | X_1, \ldots, X_{j-1} \\
& \Rightarrow X_k \perp\!\!\!\perp X_j | X_1, \ldots, X_{j-1} \\
& \Rightarrow k \text{ is not a descendant of } j \text{ in } D.
\end{aligned}$$

In the second to last line we used that $X_j = \sum_{j' \in \mathrm{pa}_D(j)} f_{j,j'}(X_{j'}) + \sigma_j Z_j$. In the last line we used faithfulness. The other direction follows from the definition of a PLSEM with DAG D: For all $j \in C_i$ we have that

$$\begin{aligned} &k \text{ is a non-descendant of } j \text{ in } D \\ &\quad \Rightarrow x_k = e_k^t F^{-1}(z) \text{ is constant in } z_j. \end{aligned} \tag{3.24}$$

By combining equation (3.23) and equation (3.24),

$$e_k^t (\mathrm{D}F)^{-1}(x) e_j = 0 \iff k \text{ is a non-descendant of } j \text{ in } D.$$

Hence, by equation (3.20) and equation (3.22)

$$\begin{aligned} &k \text{ is a non-descendant of } j \text{ in } D \text{ for all } j \in C_i \\ \iff &e_k^t (\mathrm{D}F)^{-1}_{\bullet,(i+1):p}(e_j)_{(i+1):p} \equiv 0 \text{ for all } j \in C_i \\ \iff &e_k^t (\mathrm{D}F)^{-1} \partial_i^2 F \equiv 0. \end{aligned}$$

This concludes the proof of (c).

Now let us turn to the proof of (d). Fix a DAG D'. Node k is a descendant of i in D' if and only if for all causal orderings σ of D' we have $\sigma(i) < \sigma(k)$. Hence,

$$\begin{aligned} &k \text{ is a descendant of } i \text{ in all DAGs } D' \text{ of PLSEMs that generate } \mathbb{P} \\ \iff &\text{for all causal orderings } \sigma \text{ of a DAG } D' \text{ of a PLSEM} \\ &\text{that generates } \mathbb{P}\text{: } \sigma(i) < \sigma(k). \end{aligned} \tag{3.25}$$

By Theorem 7, a permutation σ is a causal ordering of a DAG D' of a PLSEM that generates $\mathbb{P}$ if and only if $\sigma(l) < \sigma(m)$ for all $(l, m) \in \mathcal{V}$:

$$\begin{aligned} &\text{for all causal orderings } \sigma \text{ of a DAG } D' \text{ of a PLSEM} \\ &\text{that generates } \mathbb{P}\text{: } \sigma(i) < \sigma(k) \\ \iff &\text{for all permutations } \sigma \text{ with } \sigma(l) < \sigma(m) \text{ for all } (l,m) \in \mathcal{V} \\ &\text{we have } \sigma(i) < \sigma(k) \\ \iff &(i,k) \in \tilde{\mathcal{V}}. \end{aligned} \tag{3.26}$$

Combining equations (3.25) and (3.26) concludes the proof of (d). Statement (b) follows from (d), as $(i, j) \in \mathcal{V}$ implies that $(i, j) \in \tilde{\mathcal{V}}$. This concludes the proof of (b). □

3.A.6 Proofs of consistency and correctness of estimation procedures

Proof of Theorem 8

Proof. We prove that for n sufficiently large and with high probability, for any DAG $D^0 \in \mathscr{D}(\mathbb{P})$ and $D \in \mathscr{C}(D^0)$ such that (without loss of generality) D^0 and D only differ by reversal of a covered edge between nodes i and j,

$$\log(\hat{\sigma}_i^D) + \log(\hat{\sigma}_j^D) - \log(\hat{\sigma}_i^{D^0}) - \log(\hat{\sigma}_j^{D^0}) \geq 3\xi_0/4, \tag{3.27}$$

where $\hat{\sigma}_j^D$ denotes the unpenalized maximum likelihood estimator of the standard deviation of the residuals at node j in DAG D. Similarly, for n sufficiently large and with high probability, for $D^0, \tilde{D}^0 \in \mathscr{D}(\mathbb{P})$ that only differ by a reversal of a covered linear edge between nodes i and j,

$$\left|\log(\hat{\sigma}_i^{D^0}) + \log(\hat{\sigma}_j^{D^0}) - \log(\hat{\sigma}_i^{\tilde{D}^0}) - \log(\hat{\sigma}_j^{\tilde{D}^0})\right| \leq \xi_0/4. \tag{3.28}$$

The uniform bounds (3.27) and (3.28) imply that for $\alpha \in (\xi_0/4, 3\xi_0/4)$, each score-based decision whether a covered edge $i \to j$ is linear or not in step 6 of Algorithm 2 is consistent. The consistency of the estimated distribution equivalence class then follows from the correctness of Algorithm 1, which is justified at the beginning of Section 3.3.1. Obviously, the constants in (3.27) and (3.28) can be changed allowing for any $\alpha \in (0, \xi_0)$.

Proof of (3.27). By Assumption 1 (i), all DAGs under consideration have uniformly bounded node degrees. It now follows exactly along the lines of Sections 2.1, 2.2, 2.3 and 5 in the supplement to Bühlmann et al. (2014) that for $D \in \mathscr{D}(\mathbb{P}) \cup \mathscr{C}(\mathscr{D}(\mathbb{P}))$,

$$(\sigma_k^D)^2 \leq (\hat{\sigma}_k^D)^2 + \Delta_{n,k}^D, \tag{3.29}$$

and for $D^0 \in \mathscr{D}(\mathbb{P})$,

$$(\hat{\sigma}_k^{D^0})^2 \leq (\sigma_k^{D^0})^2 + \gamma_{n,k}^{D^0} + \Delta_{n,k}^{D^0}, \tag{3.30}$$

with $\Delta_{n,k}^D$, $\gamma_{n,k}^{D^0}$ as defined in Assumptions 1 (iii) and (iv).

Without loss of generality, let $D^0 \in \mathscr{D}(\mathbb{P})$ and $D \in \mathscr{C}(D^0)$ such that D^0 and D only differ by reversal of a covered nonlinear edge between nodes i

and j. Substituting (3.29) and (3.30) and then using (3.9) and (3.10),

$$\sum_{k\in\{i,j\}} \left(\log(\hat{\sigma}_k^D) - \log(\hat{\sigma}_k^{D^0})\right) \geq \sum_{k\in\{i,j\}} \left(\log(\sigma_k^D) - \log(\sigma_k^{D^0})\right) + R_{n,D,D^0}$$
$$\geq \xi_p + R_{n,D,D^0},$$

where

$$R_{n,D,D^0} = \frac{1}{2} \sum_{k\in\{i,j\}} \left(\log\left(1 + \frac{-\Delta_{n,k}^D}{(\sigma_k^D)^2}\right) - \log\left(1 + \frac{\gamma_{n,k}^{D^0} + \Delta_{n,k}^{D^0}}{(\sigma_k^{D^0})^2}\right)\right).$$

By Assumption 1 (ii), the error variances are bounded away from zero. Using Taylor expansion and Assumptions 1 (iii) and (iv), $R_{n,D,D^0} = o_P(1)$. As ξ_p is uniformly bounded from below by ξ_0, we have that

$$\sum_{k\in\{i,j\}} \left(\log(\hat{\sigma}_k^D) - \log(\hat{\sigma}_k^{D^0})\right) \geq \xi_0 + o_P(1).$$

Therefore, for n sufficiently large and with high probability,

$$\sum_{k\in\{i,j\}} \left(\log(\hat{\sigma}_k^D) - \log(\hat{\sigma}_k^{D^0})\right) \geq 3\xi_0/4.$$

A completely analogous argument yields (3.28) for $D^0, \tilde{D}^0 \in \mathscr{D}(\mathbb{P})$. □

Proof of Lemma 5

Proof. (a): For a DAG D with $i \to j$ in D, we denote by $D_{i\leftarrow j}$ the graph that differs from D only by the reversal of $i \to j$. Let $i \to j$ in $\mathcal{K}$. Recall that $G_{P,\mathcal{K}}$ is obtained by imposing all edge orientations in $\mathcal{K}$ on the pattern P and closing orientations under R1-R4 in Figure 3.6. Hence, by construction, $i \to j$ in $G_{P,\mathcal{K}}$. Throughout the proof we use that for a background knowledge $\mathcal{K}'$, by Theorems 2 and 4 in Meek (1995), the set of all consistent DAG extensions of $G_{P,\mathcal{K}'}$ equals the set of all Markov equivalent DAGs that have pattern P and edge orientations that comply with the background knowledge $\mathcal{K}'$. Then, it holds that

$\exists$ consistent DAG extension D of $G_{P,\mathcal{K}}$ in which $i \to j$ is covered
$\iff$ $\exists$ consistent DAG extension D of $G_{P,\mathcal{K}}$: $D_{i\leftarrow j}$ is Markov equivalent to D
$\iff$ $\exists$ consistent DAG extension D of $G_{P,\mathcal{K}\setminus\{i\to j\}}$: $D_{i\leftarrow j}$ is Markov equivalent to D,

where the first equivalence follows from Lemma 1 in Chickering (1995). By assumption, $i \to j$ is not in P and not in the background knowledge. Hence, the soundness of the four orientation rules R1-R4 (Meek, 1995, Theorem 2) implies that $i - j$ in $G_{P,\mathcal{K}\setminus\{i\to j\}}$ which is the case if and only if $G_{P,\mathcal{K}} \neq G_{P,\mathcal{K}\setminus\{i\to j\}}$.
To show the other implication, by the completeness of the orientation rules (Meek, 1995, Theorem 4),

$$\begin{aligned} & i - j \text{ in } G_{P,\mathcal{K}\setminus\{i\to j\}} \\ \Longrightarrow\ & \exists \text{ consistent DAG extensions } D_1, D_2 \text{ of } G_{P,\mathcal{K}\setminus\{i\to j\}}\text{:} \\ & i \to j \text{ in } D_1 \text{ and } i \leftarrow j \text{ in } D_2 \\ \Longrightarrow\ & \exists \text{ consistent DAG extension } D \text{ of } G_{P,\mathcal{K}\setminus\{i\to j\}}\text{:} \\ & D_{i\leftarrow j} \text{ is a consistent DAG extension of } G_{P,\mathcal{K}\setminus\{i\to j\}}, \end{aligned}$$

where the last implication follows from Theorem 2 in Chickering (1995). As both, D and $D_{i\leftarrow j}$ are consistent DAG extension of the same pattern P, $D_{i\leftarrow j}$ is Markov equivalent to D, which finishes the proof.

(b) As $G_{P,\mathcal{K}} \neq G_{P,\mathcal{K}\setminus\{i\to j\}}$, by Lemma 5 (a), there exists a consistent DAG extension of $G_{P,\mathcal{K}}$ in which $i \to j$ is covered. From that, it immediately follows that all $k \in \mathrm{pa}_{G_{P,\mathcal{K}}}(i)$ are adjacent to j in $G_{P,\mathcal{K}}$. As $G_{P,\mathcal{K}}$ is closed under R1-R4, $k \to j$ in $G_{P,\mathcal{K}}$ due to R2 and $\mathrm{pa}_{G_{P,\mathcal{K}}}(i) \subseteq \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$. Analogously, for $k' \in \mathrm{pa}_{G_{P,\mathcal{K}}}(j) \setminus \{i\}$, either $k' \to i$ or $k' - i$ in $G_{P,\mathcal{K}}$ as $i \to j$ can be covered.

Step 1: Orient $k' - i$ with $k' \in \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$ into i. To be precise, define the new background knowledge $\widetilde{\mathcal{K}} := \mathcal{K} \cup (\bigcup_{k' \in \mathrm{pa}_{G_{P,\mathcal{K}}}(j)} \{k' \to i\})$. As there is a consistent DAG extension D of $G_{P,\mathcal{K}}$ in which $i \to j$ is covered, $k' \to i$ in D for all $k' \in \mathrm{pa}_{G_{P,\mathcal{K}}}(j) \setminus \{i\}$. Hence, by definition, $\widetilde{\mathcal{K}}$ is consistent.

Step 2: Close orientations under R1-R4 to obtain the maximally oriented PDAG $G_{P,\widetilde{\mathcal{K}}}$ with respect to the pattern P and background knowledge $\widetilde{\mathcal{K}}$ (Meek, 1995).

Claim 1: Let (x, y, z), $z \in \{i, j\}$ be a triple such that $x - y$ or $y - z$ in $G_{P,\mathcal{K}}$ and $x \to y \to z$ in $G_{P,\widetilde{\mathcal{K}}}$. Then $y = i, z = j$ or $x \to y - z$ in $G_{P,\mathcal{K}}$.
Suppose that $y \neq i$ and $x - y$ in $G_{P,\mathcal{K}}$. Then, $x \to y$ in $G_{P,\widetilde{\mathcal{K}}}$ was oriented in Step 2 by applying R1-R4. We will lead this to a contradiction. By Lemma 13, y is a descendant of i in $G_{P,\widetilde{\mathcal{K}}}$. Moreover, recall that $y \to z$ in $G_{P,\widetilde{\mathcal{K}}}$. By construction, there exists a consistent DAG extension D of $G_{P,\widetilde{\mathcal{K}}}$

in which $i \to j$ is covered. As $y \neq i$ and $z \in \{i, j\}$, $y \to i$ and $y \to j$ in D. But y is a descendant of i in D, so D contains a cycle. Contradiction.

Claim 2: In Step 2*, no additional edges are oriented into i or j.*
Suppose the contrary and consider the first edge that is oriented into i or j by applying one of R1-R4 in Step 2. We will consider the rules case-by-case.

R1: Let x be the upper, y the lower left and z the right node in R1 in Figure 3.6. Suppose $y \to z$, $z \in \{i, j\}$ is implied by R1 in Step 2. Thus, $x \to y \to z$ in $G_{P,\widetilde{\mathcal{K}}}$ and $y - z$ in $G_{P,\mathcal{K}}$. As $G_{P,\mathcal{K}}$ is closed under R1-R4, $x - y$ in $G_{P,\mathcal{K}}$. Then, by *Claim 1*, $y = i$ and $z = j$. But $i \to j$ in $G_{P,\mathcal{K}}$, contradiction.

R2: Let x be the left, y the middle and z the right node in R2 in Figure 3.6. Thus, $x \to y \to z$ with $x \to z$ in $G_{P,\widetilde{\mathcal{K}}}$, and $x - z$ in $G_{P,\mathcal{K}}$.
Case 1: $y = i$. As $z \in \{i, j\}$, $z = j$. As $i \to j$ and $x - z = j$ in $G_{P,\mathcal{K}}$ and $G_{P,\mathcal{K}}$ is closed under R1-R4, $x - y = i$ in $G_{P,\mathcal{K}}$. Therefore, $x \to y = i$ in $G_{P,\widetilde{\mathcal{K}}}$ was either oriented in Step 1 or Step 2. It was not oriented in Step 1 as $x - z = j$ (that is, $x \notin \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$). Also, it was not oriented in Step 2 as by assumption, $x \to z$ is the first edge oriented into i or j in Step 2. Contradiction.
Case 2: $y \neq i$. By *Claim 1*, $x \to y - z$ in $G_{P,\mathcal{K}}$. By assumption, $x - z$ is the first edge that is oriented into $z \in \{i, j\}$ in Step 2. Hence, $y - z$ was oriented into z in Step 1 ($y \in \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$) and $x - z$ not ($x \notin \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$). As x is oriented into $z \in \{i, j\}$ in $G_{P,\widetilde{\mathcal{K}}}$ and as there is a consistent DAG extension of $G_{P,\widetilde{\mathcal{K}}}$ in which $i \to j$ is covered, $x - i$ and $x - j$ in $G_{P,\mathcal{K}}$. Recall that $x \to y$ in $G_{P,\mathcal{K}}$ and $y \in \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$. As $G_{P,\mathcal{K}}$ is closed under R1-R4, $x \to j$ in $G_{P,\mathcal{K}}$ by R2, which contradicts $x \notin \mathrm{pa}_{G_{P,\mathcal{K}}}(j)$.

R3: Does not apply. Step 1 and Step 2 do not create new v-structures.

R4: Let x be the upper left, y the upper right and z the lower right node in R4 in Figure 3.6. We have $x \to y \to z$ in $G_{P,\widetilde{\mathcal{K}}}$ and $x - y$ or $y - z$ in $G_{P,\mathcal{K}}$. Note that x and z are not adjacent in $G_{P,\mathcal{K}}$ and $G_{P,\widetilde{\mathcal{K}}}$. If $y = i$, then $z = j$ and x is a parent of i but not adjacent to j in $G_{P,\widetilde{\mathcal{K}}}$. This contradicts the fact that $i \to j$ is covered in a consistent DAG extension of $G_{P,\widetilde{\mathcal{K}}}$. Hence, $y \neq i$. By *Claim 1*, $x \to y - z$ in $G_{P,\mathcal{K}}$. As x is not adjacent to z in $G_{P,\mathcal{K}}$, $y \to z$ in $G_{P,\mathcal{K}}$ by R1, which contradicts the fact that $G_{P,\mathcal{K}}$ is closed under R1-R4. This concludes the proof of *Claim 2*.

By *Claim 2*, we do not orient edges into i or j in Step 2. Hence, by construction, $G_{P,\widetilde{\mathcal{K}}}$ satisfies $\mathrm{pa}_{G_{P,\widetilde{\mathcal{K}}}}(i) = \mathrm{pa}_{G_{P,\widetilde{\mathcal{K}}}}(j) \setminus \{i\} = \mathrm{pa}_{G_{P,\mathcal{K}}}(j) \setminus \{i\}$.

By Lemma 14, there exists a consistent DAG extension D of $G_{P,\widetilde{\mathcal{K}}}$ in which all undirected edges incident to i and j are oriented out of i and j. By construction, D is a consistent DAG extension of $G_{P,\mathcal{K}}$. Moreover, $\mathrm{pa}_D(i) = \mathrm{pa}_D(j) \setminus \{i\} = \mathrm{pa}_{G_{P,\mathcal{K}}}(j) \setminus \{i\}$, which concludes the proof. □

Lemma 13. *Consider the maximally oriented PDAG $G_{P,\mathcal{K}'}$ (with orientations closed under R1-R4) with respect to the pattern P and consistent background knowledge $\mathcal{K}'$. Let $a_m - b$ in $G_{P,\mathcal{K}'}$ for all $1 \leq m \leq M$ and assume there exists a consistent DAG extension of $G_{P,\mathcal{K}'}$ in which $a_m \to b$ for all $1 \leq m \leq M$. We orient $a_m \to b$ for all $m \leq M$ and close the orientations under R1-R4. Let us denote the edges we orient $a'_m \to b'_m$, $m = 1, 2, \ldots$. Then $b'_m, m \geq 0$ are descendants of b.*

Proof. By induction. By definition, b is a descendant of b. At each step, apply one of R1-R4 and orient $a'_m \to b'_m$. This only occurs if one of the directed edges in one of R1-R4 is actually an edge $a_k \to b$ or $a'_k \to b'_k$ that was oriented at an earlier stage $1 \leq k \leq M$ (in the first case) or $k < m$ (in the second case). By the induction assumption, b'_k is a descendant of b for all $k < m$. By looking at Figure 3.6 (i.e. going through the cases R1-R4) we can see that in each case, b'_m is a descendant of b (in the first case) or b'_k (in the second case). Hence b'_m is a descendant of b. □

Lemma 14. *Consider the maximally oriented PDAG $G_{P,\mathcal{K}}$ (with orientations closed under R1-R4) with respect to the pattern P and consistent background knowledge $\mathcal{K}$. Let $x \to y$ in $G_{P,\mathcal{K}}$. Then, there exists a consistent DAG extension of $G_{P,\mathcal{K}}$ in which all undirected edges incident to x and y are oriented out of x and y.*

Proof. Orient an undirected edge e incident to y out of y and close the orientations under R1-R4. By Theorems 2 and 4 in Meek (1995) the resulting PDAG is maximally oriented with respect to the pattern P and consistent background knowledge $K \cup \{e\}$. By Lemma 13, no edge that is oriented in that process will point into x or y. Now repeat, until there is no more undirected edge incident to y. Then, analogously orient all undirected edges incident to x out of x. □

Proof of Lemma 6

Proof. We first prove that for $k \geq 1$, by construction, each $G_{P,\mathcal{K}_k}$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$. This means that $G_{P,\mathcal{K}_k}$ and $G_{\mathscr{D}(\mathbb{P})}$ have

the same skeleton and v-structures and $i \to j$ in $G_{\mathscr{D}(\mathbb{P})} \Rightarrow i \to j$ in $G_{P,\mathcal{K}_k}$. Then, we show that there exists a $k_0 \leq |\mathcal{K}_1^{\text{init}}|+1$ such that either $\mathcal{K}_{k_0}^{\text{init}} = \emptyset$ or there is no edge in $\mathcal{K}_{k_0}^{\text{init}}$ that is covered in any of the consistent DAG extensions of $G_{P,\mathcal{K}_{k_0}}$. For k_0 it holds that $G_{P,\mathcal{K}_{k_0}} = G_{\mathscr{D}(\mathbb{P})}$.
By construction, $G_{P,\mathcal{K}_1} = D^0$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$. By Theorem 5 (b), if $\mathcal{K}_1^{\text{init}} = \emptyset$ or none of the edges in $\mathcal{K}_1^{\text{init}}$ are covered in D^0, $G_{\mathscr{D}(\mathbb{P})} = D^0 = G_{P,\mathcal{K}_1}$. For a fixed $k \geq 1$, suppose that $G_{P,\mathcal{K}_k}$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$ and that there exists $\{i \to j\} \in \mathcal{K}_k^{\text{init}}$ that is covered in a consistent DAG extension of $G_{P,\mathcal{K}_k}$, which we denote by D. By assumption, as $G_{P,\mathcal{K}_k}$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$, D is a consistent DAG extension of $G_{\mathscr{D}(\mathbb{P})}$. Hence, $D \in \mathscr{D}(\mathbb{P})$ by Theorem 4.
Case 1: As $i \to j$ is covered and linear in $D \in \mathscr{D}(\mathbb{P})$, by Theorem 5 (a), there is a DAG $D' \in \mathscr{D}(\mathbb{P})$ with $i \leftarrow j$. Therefore, by Definition 1, $i - j$ in $G_{\mathscr{D}(\mathbb{P})}$.
By construction, $\mathcal{K}_{k+1} = \mathcal{K}_k \setminus \{i \to j\}$. Hence, by Lemma 5 (a), $G_{P,\mathcal{K}_{k+1}}$ equals $G_{P,\mathcal{K}_k}$ except for an undirected edge $i - j$ (all other directed edges in $G_{P,\mathcal{K}_k}$ are either directed in P or still contained in $\mathcal{K}_{k+1}$, hence they must be directed in $G_{P,\mathcal{K}_{k+1}}$). Therefore, $G_{P,\mathcal{K}_{k+1}}$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$.
Case 2: As $i \to j$ is covered and nonlinear in $D \in \mathscr{D}(\mathbb{P})$, by Theorem 5 (a), $i \to j$ in all DAGs in $\mathscr{D}(\mathbb{P})$. Hence, $i \to j$ in $G_{\mathscr{D}(\mathbb{P})}$ by Definition 1.
By construction, $\mathcal{K}_{k+1} = \mathcal{K}_k$ and $G_{P,\mathcal{K}_{k+1}} = G_{P,\mathcal{K}_k}$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$. Moreover, as $\{i \to j\} \notin \mathcal{K}_{k+1}^{\text{init}}$ and $\{i \to j\} \in \mathcal{K}_{k+1}^{\text{nonl}}$, $i \to j$ is fixed in all $G_{P,\mathcal{K}_l}$ for $l > k$.

In both cases, $|\mathcal{K}_{k+1}^{\text{init}}| = |\mathcal{K}_k^{\text{init}}| - 1$. Hence, there exists a $k_0 \leq |\mathcal{K}_1^{\text{init}}| + 1$ such that either $\mathcal{K}_{k_0}^{\text{init}} = \emptyset$ or no edge in $\mathcal{K}_{k_0}^{\text{init}}$ is covered in any of the consistent DAG extensions of $G_{P,\mathcal{K}_{k_0}}$. We will now prove that $G_{P,\mathcal{K}_{k_0}} = G_{\mathscr{D}(\mathbb{P})}$. If $\mathcal{K}_{k_0}^{\text{init}} = \emptyset$, this immediately follows from Case 1 and Case 2. For $\mathcal{K}_{k_0}^{\text{init}} \neq \emptyset$, suppose $G_{P,\mathcal{K}_{k_0}} \neq G_{\mathscr{D}(\mathbb{P})}$. Then, there are $M \geq 1$ undirected edges $i_m - j_m, m = 1, ..., M$, in $G_{\mathscr{D}(\mathbb{P})}$ with $i_m \to j_m$ in $G_{P,\mathcal{K}_{k_0}}$. By construction, $\{i_m \to j_m\}_{m=1,...,M} \subseteq \mathcal{K}_{k_0}$. From Case 2 it must hold that $\{i_m \to j_m\}_{m=1,...,M} \subseteq \mathcal{K}_{k_0}^{\text{init}}$. By Theorem 5 (b), $\mathscr{D}(\mathbb{P})$ is connected with respect to covered linear edge reversals. Hence, there is an $1 \leq m_0 \leq M$ for which $i_{m_0} \to j_{m_0}$ is covered in a consistent DAG extension of $G_{P,\mathcal{K}_{k_0}}$. But $\{i_{m_0} \to j_{m_0}\} \in \mathcal{K}_{k_0}^{\text{init}}$, contradiction. We just showed that $G_{\mathscr{D}(\mathbb{P})}$ is a consistent extension of $G_{P,\mathcal{K}_{k_0}}$. As by construction, $G_{P,\mathcal{K}_{k_0}}$ is a consistent extension of $G_{\mathscr{D}(\mathbb{P})}$, we conclude that $G_{P,\mathcal{K}_{k_0}} = G_{\mathscr{D}(\mathbb{P})}$, which finishes the proof. □

Chapter 4

Estimation of total causal effects in nonparametric models[1]

We consider the problem of inferring the total causal effect of a single continuous variable intervention on a (response) variable of interest. We propose a certain marginal integration regression technique for a very general class of potentially nonlinear structural equation models (SEMs) with known structure, or at least known superset of adjustment variables: we call the procedure S-mint *regression. We easily derive that it achieves the convergence rate as for nonparametric regression: for example, single variable intervention effects can be estimated with convergence rate* $n^{-2/5}$ *assuming smoothness with twice differentiable functions. Our result can also be seen as a major robustness property with respect to model misspecification which goes much beyond the notion of double robustness. When the structure of the SEM is not known, we can estimate (the equivalence class of) the directed acyclic graph corresponding to the SEM, and then proceed by using* S-mint *based on these estimates. We empirically compare the* S-mint *regression method with more classical approaches and argue that the former is indeed more robust, more reliable and substantially simpler.*

[1]This chapter is a slightly modified version of the published article Ernest, J. and Bühlmann, P. (2015). „Marginal integration for nonparametric causal inference“. *Electronic Journal of Statistics* 9 (2), pp. 3155–3194. DOI: 10.1214/15-EJS1075.

4.1 Introduction

Understanding cause-effect relationships between variables is of great interest in many fields of science. An ambitious but highly desirable goal is to infer causal effects from observational data obtained by observing a system of interest without subjecting it to interventions.[2] This would allow to circumvent potentially severe experimental constraints or to substantially lower experimental costs. The words "causal inference" (usually) refer to the problem of inferring effects which are due to (or caused by) interventions: if we make an outside intervention at a variable X, say, what is its effect on another response variable of interest Y. We describe examples in Section 4.1.3. Various fields and concepts have contributed to the understanding and quantification of causal inference: the framework of potential outcomes and counterfactuals (cf. Rubin, 2005), see also Dawid (2000), structural equation modeling (cf. Bollen, 1998), and graphical modeling (cf. Greenland et al., 1999; Lauritzen and Spiegelhalter, 1988); the book by Pearl (2000) provides a nice overview.

We consider aspects of the problem indicated above, namely inferring intervention or causal effects from observational data without external interventions. Thus, we deal (in part) with the question of how to infer causal effects *without* relying on randomized experiments or randomized studies. Besides fundamental conceptual aspects, as treated for example in the books by Pearl (2000), Spirtes et al. (2000) and Koller and Friedman (2009), important issues include statistical tasks such as estimation accuracy and robustness with respect to model misspecification. This work focuses on the two latter topics, covering also high-dimensional sparse settings with many variables (parameters) but relatively few observational data points.

In general, the tools for inferring causal effects are different from regression methods, but as we will argue, the regression methods, when properly applied, remain a useful tool for causal inference. In fact, for the estimation of total causal effects, we make use of a marginal integration regression method which has originally been proposed for additive regression modeling (Linton and Nielsen, 1995). Its use in causal inference is novel. Relying on known theory for marginal integration in regression (Fan et al., 1998), our main result (Theorem 12) establishes optimal convergence properties

[2] More generally, in the presence of both, interventional and observational data, the goal is to infer intervention or causal effects among variables which are not directly targeted by the interventions from interventional data.

and justifies the method as a fully robust procedure against model misspecification, as explained further in Section 4.1.2.

4.1.1 Basic concepts and definitions for the estimation of causal effects under interventions

We very briefly introduce some of the basic concepts for the estimation of causal effects under interventions. We consider p random variables $X_1, \ldots, X_p$, where one of them is a response variable Y of interest and one of them an intervention variable X, that is, the variable where we make an external intervention by setting X to a certain value x. Such an intervention is denoted by Pearl's do-operator $\mathrm{do}(X = x)$ (cf. Pearl, 2000). We denote the indices corresponding to Y and X by j_Y and j_X, respectively: thus, $Y = X_{j_Y}$ and $X = X_{j_X}$. We assume a setting where all relevant variables are observed, that is, there are no relevant hidden variables.[3]

The system of variables is assumed to be generated from a structural equation model (SEM):

$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j),\ j = 1, \ldots, p. \tag{4.1}$$

Thereby, $\varepsilon_1, \ldots, \varepsilon_p$ are independent noise (or innovation) variables, and there is an underlying structure given in terms of a directed acyclic graph (DAG) D, where each node j corresponds to the random variable X_j: We denote by $\mathrm{pa}(j) = \mathrm{pa}_D(j)$ the set of parents of node j in the underlying DAG D,[4] and $f_j(\cdot)$ are assumed to be real-valued (measurable) functions. For any index set $U \subseteq \{1, \ldots, p\}$ we write $X_U := (X_v)_{v \in U}$, for example, $X_{\mathrm{pa}(j)} = (X_v)_{v \in \mathrm{pa}(j)}$.

The causal mechanism we are interested in is the total effect of an intervention at a single variable X on a response variable Y of interest.[5] The distribution of Y when doing an external intervention $\mathrm{do}(X = x)$ by setting variable X to x is identified with its density (assumed to exist) or discrete probability function and is denoted by $p(y|\mathrm{do}(X = x))$. The mathematical definition of $p(y|\mathrm{do}(X = x))$ can be given in terms of a so-called truncated Markov factorization or maybe more intuitively, by direct plug-in of the intervention value x for variable X and propagating

[3] It suffices to assume that Y, X and $X_{\mathrm{pa}(j_X)}$ (the parents of X) are observed, see (4.3).
[4] The set of parents is $\mathrm{pa}_D(j) = \{k;$ there exists a directed edge $k \to j$ in DAG $D\}$.
[5] A total effect is the effect of an intervention at a variable X to another variable Y, taking into account the total of all (directed) paths from X to Y.

this intervention value x to all other random variables including Y in the structural equation model (4.1); precise definitions are, for example, given in Pearl (2000) or Spirtes et al. (2000). The underlying important assumption in the definition of $p(y|\mathrm{do}(X = x))$ is that the functional forms and error distributions of the structural equations for all the variables X_j which are different from X do not change when making an intervention at X.

A very powerful representation of the intervention distribution is given by the well-known backdoor adjustment formula.[6] We say that a path in a DAG D is blocked by a set of nodes S if and only if it contains a chain $.. \to m \to ..$ or a fork $.. \leftarrow m \to ..$ with $m \in S$ or a collider $.. \to m \leftarrow ..$ such that $m \notin S$ and no descendant of m is in S. Furthermore, a set of variables S is said to satisfy the backdoor criterion relative to (X, Y) if no node in S is a descendant of X and if S blocks every path between X and Y with an arrow pointing into X. For a set S that satisfies the backdoor criterion relative to (X, Y), the backdoor adjustment formula reads:

$$p(y|\mathrm{do}(X = x)) = \int p(y|X = x, X_S)dP(X_S), \tag{4.2}$$

where $p(\cdot)$ and $P(\cdot)$ are generic notations for the density or distribution, respectively (Pearl, 2000, Theorem 3.3.2). An important special case of the backdoor adjustment formula is obtained when considering the adjustment set $S = \mathrm{pa}(j_X)$: if $j_Y \notin \mathrm{pa}(j_X)$, that is, if Y is not in the parental set of the variable X, then:

$$p(y|\mathrm{do}(X = x)) = \int p(y|X = x, X_{\mathrm{pa}(j_X)})dP(X_{\mathrm{pa}(j_X)}). \tag{4.3}$$

Thus, if the parental set $\mathrm{pa}(j_X)$ is known, the intervention distribution can be calculated from the standard observational conditional and marginal distributions. Our main focus is the expectation of Y when doing the intervention $\mathrm{do}(X = x)$, the so-called total effect:

$$\mathbb{E}[Y|\mathrm{do}(X = x)] = \int y\, p(y|\mathrm{do}(X = x))dy.$$

A general and often used route for inferring $\mathbb{E}[Y|\mathrm{do}(X = x)]$ is as follows: the directed acyclic graph (DAG) corresponding to the structural equation

[6] For a simple version of the formula, skip the text until the second line after formula (4.2).

model (SEM) is either known or (its Markov equivalence class) estimated from data; building on this, one can estimate the functions in the SEM (edge functions in the DAG), the error distributions in the SEM, and finally extract an estimate of $\mathbb{E}[Y|\mathrm{do}(X = x)]$ (or bounds of this quantity if the DAG is not identifiable) from the observational distribution. See, for example, Maathuis et al. (2009), Pearl (2000), Spirtes (2010), and Spirtes et al. (2000).

4.1.2 Our contribution

The new results from this chapter should be explained for two different scenarios and application areas: one where the structure of the DAG D in the SEM is known, and the other where the structure and the DAG D are unknown and estimated from data. Of course, the second setting is linked to the first by treating the estimated as the true known structure. However, due to estimation errors, a separate discussion is in place.

Structural equation models with known structure

We consider a general SEM as in (4.1) with known structure in form of a DAG D but unknown functions f_j and unknown error distributions for ε_j. As already mentioned before, our focus is on inferring the total effect

$$\mathbb{E}[Y|\mathrm{do}(X = x)] = \int y\, p(y|\mathrm{do}(X = x))dy, \tag{4.4}$$

where $p(y|\mathrm{do}(X = x))$ is the interventional density (or discrete probability function) of Y as loosely described in Section 4.1.1.

The first approach to infer the total effect in (4.4) is to estimate the functions f_j and error distributions for ε_j in the SEM. It is then possible to calculate $\mathbb{E}[Y|\mathrm{do}(X = x)]$, typically using a path-based method based on the DAG D (see also Section 4.3.1). This route is essentially impossible without putting further assumptions on the functional form of f_j in the SEM (4.1). For example, one often makes the assumption of additive errors, and if the cardinality of the parental set $|\mathrm{pa}(j)|$ is large, additional constraints like additivity of a nonparametric function are in place to avoid the curse of dimensionality. Thus, by keeping the general possibly non-additive structure of the functions f_j in the SEM, we have to reject this approach.

The second approach for inferring the total effect in (4.4) relies on the powerful backdoor adjustment formula in (4.2). At first sight, the problem seems ill-posed because of the appearance of $p(Y|X = x, X_S)$ for a set S with possibly large cardinality $|S|$. But since we integrate over the variables X_S in (4.2), we are *not* entering the curse of dimensionality. This simple observation is a key idea of this work. We present an estimation technique for $\mathbb{E}[Y|\mathrm{do}(X = x)]$, or other functionals of $p(y|\mathrm{do}(X = x))$, using marginal integration which has been proposed and analyzed for additive regression modeling (Linton and Nielsen, 1995). The idea of our marginal integration approach is to first estimate a fully nonparametric regression of Y versus X and the variables X_S from a valid adjustment set satisfying the backdoor criterion (for example the parents of X or a superset thereof) and then average the obtained estimate over the variables X_S. We call the procedure "*S-mint*" standing for *m*arginal *int*egration with adjustment set S.

Our main result in Theorem 12 establishes that $\mathbb{E}[Y|\mathrm{do}(X = x)]$ can be inferred via marginal integration with the same rate of convergence as for one-dimensional nonparametric function estimation for a very large class of structural equation models with potentially non-additive functional forms in the equations. We thereby achieve a major robustness result against model misspecification, as we only assume some standard smoothness assumptions but no further conditions on the functional form or nonlinearity of the functions f_j in the SEM, not even additive errors. Our main result (Theorem 12) also applies using a superset of the true underlying DAG D (i.e., there might be additional directed edges in the superset), see Section 4.2.3. For example, such a superset could arise from knowing the order of the variables (e.g., in a time series context), or an approximate superset might be available from estimation of the DAG where one would not care too much about slight or moderate overfitting.

Inferring $\mathbb{E}[Y|\mathrm{do}(X = x)]$ under model-misspecification is the theme of double robustness in causal inference, typically with a binary treatment variable X (cf. van der Laan and Robins, 2003). There, misspecification of either the regression or the propensity score model[7] is allowed but at least one of them has to be correct to allow for consistent estimation: the terminology "double robustness" is intended to reflect this kind of robustness. In contrast to double robustness, we achieve here "full robustness" where essentially any form of "misspecification" is allowed, in the sense that *S-mint* does not require any specification of the functional form of the

[7]Definitions can be found in Section 4.2.1

structural equations in the SEM. More details are given in Section 4.2.1.

The local nature of parental sets. Our *S-mint* procedure requires the specification of a valid adjustment set S: as described in (4.3), we can always use the parental set $\text{pa}(j_X)$ if $j_Y \notin \text{pa}(j_X)$. The parental variables are often an interesting choice for an adjustment set which corresponds to a *local* operation. Furthermore, as discussed below, the local nature of the parental sets can be very beneficial in presence of only approximate knowledge of the true underlying DAG D.

Structural equation models with unknown structure

Consider the SEM (4.1), but now we assume that the DAG D is unknown. For this setting, we propose a two-stage scheme ("*est S-mint*", see Section 4.3.5). First, we estimate the structure of the DAG (or the Markov equivalence class of DAGs) or the order of the variables from observational data. To do this, all of the current approaches make further assumptions for the SEM in (4.1), see, for example, Bühlmann et al. (2014), Chickering (2002), Hoyer et al. (2009), Kalisch and Bühlmann (2007), Schmidt et al. (2007), Shimizu et al. (2006), Shojaie and Michailidis (2010), and Teyssier and Koller (2005).

We can then infer $\mathbb{E}[Y|\text{do}(X = x)]$ as before with *S-mint* model fitting, but based on an estimated (instead of the true) adjustment set S. This seems often more advisable than using the estimated functions in the SEM, which are readily available from structure estimation, and pursuing a path-based method with the estimated DAG. Since estimation of (the Markov equivalence class of) the DAG or of the order of variables is often very difficult and with limited accuracy for finite sample size, the second stage with *S-mint* model fitting is fairly robust with respect to errors in order- or structure-estimation and model misspecification, as suggested by our empirical results in Section 4.5.3. Therefore, such a two-stage procedure with structure- or order-search[8] and subsequent marginal integration leads to reasonably accurate and sometimes better results. For example, Section 4.5 reports a comparable performance to the direct CAM method from Section 2.5 with subsequent path-based estimation of causal effects, which is based on, or assuming, a correctly specified additive SEM. Thus, even if the *est S-mint* approach with fully nonparametric *S-mint* modeling

[8] We do not make use of, e.g., estimated edge functions, even if they were implicitly estimated for structure-search, as, for example, in Chickering, 2002.

in the second stage is not exploiting the additional structural assumption of an additive SEM, it exhibits a competitive performance.

As mentioned in the previous subsection, the parental sets (or supersets thereof) with their local nature are often a very good choice in presence of estimation errors with respect to inferring the true DAG (or equivalence class thereof): instead of assuming high accuracy for recovering the entire (equivalence class of the) DAG, we only need to have a reasonably accurate estimate of the much smaller and local parental set.

A combined structured (or parametric) and fully nonparametric approach. The two-stage *est S-mint* procedure is typically a combination of a structured nonparametric or parametric approach for estimating the DAG (or the equivalence class thereof) and the fully nonparametric *S-mint* method in the subsequent second stage. As outlined above, it exhibits comparatively good performance. One could think of pursuing the first stage in a fully nonparametric fashion as well, for example, by using the PC-algorithm with nonparametric conditional independence tests (Spirtes et al., 2000), see also Song et al. (2013). For finite amount of data and a fairly large number of variables, this is a very ambitious if not ill-posed task. In view of this, we almost have to make additional structural or parametric assumptions for structure learning of the DAG (or its equivalence class). However, since the fully nonparametric *S-mint* procedure in the second stage is less sensitive to incorrect specification of the DAG (or its equivalence class), the combined approach exhibits better robustness. Vice-versa, if the structural or parametric model is correct which is used for structural learning in the first stage, we do not lose much efficiency when "throwing away" (or not exploiting) such structural information in the second stage with *S-mint*. We only have empirical results to support such accuracy statements.

4.1.3 The scope of possible applications

Genetic network inference is a prominent example where causal inference methods are used; mainly for estimating an underlying network in terms of a directed graph (cf. Friedman, 2004; Husmeier, 2003; Smith et al., 2002; Yu et al., 2004). The goal is very ambitious, namely to recover relevant edges in a complex network from observational or a few interventional data.

This work does not address this issue: instead of recovering a network

(structure), inferring *total* causal or intervention effects from observational data is a different, maybe more realistic but still very challenging goal in its full generality. Yet making progress can be very useful in many areas of applications, notably for prioritizing and designing future randomized experiments which have a large total effect on a response variable of interest, ranging from molecular biology and bioinformatics (Editorial Nature Methods, 2010) to many other fields including economy, medicine or social sciences. Such model-driven prioritization for gene intervention experiments in molecular biology has been experimentally validated with some success (Maathuis et al., 2010; Stekhoven et al., 2012).

We will discuss an application from molecular biology on a rather "toy-like" level in Section 4.6. Despite all simplifying considerations, however, we believe that it indicates a broader scope of possible applications. When having approximate knowledge of the parental set of the variables in a potentially large-scale system, one would not need to worry much about the underlying form of the dependencies of (or structural equations linking) the variables: for quantifying the effect of single variable interventions, the proposed *S-mint* marginal integration estimator converges with the univariate rate, as stated in (the main result) Theorem 12.

Quantifying single variable interventions from observational data is indeed a useful first step. Further work is needed to address the following issues:

(i) inference in settings with additional hidden, unobserved variables (cf. Colombo et al., 2012; Shpitser et al., 2011; Spirtes et al., 2000; Zhang, 2008).

(ii) inference based on a combination of observational and interventional data (cf. Hauser and Bühlmann, 2012, 2014, 2015; He and Geng., 2008).

(iii) development of sound tools and methods towards more confirmatory conclusions.

The appropriate modifications and further developments of our new results (mainly Theorem 12) towards these points (i)-(iii) are not straightforward.

4.2 Causal effects for general nonlinear systems via backdoor adjustment: Marginal integration suffices

We present here the, maybe surprising, result that marginal integration allows us to infer the causal effect of a single variable intervention with a convergence rate as for one-dimensional nonparametric function estimation in essentially *any* nonlinear structural equation model.

We assume a structural equation model (as already introduced in Section 4.1.1)

$$X_j \leftarrow f_j^0(X_{\mathrm{pa}(j)}, \varepsilon_j), \qquad j = 1, \ldots, p, \tag{4.5}$$

where $\varepsilon_1, \ldots, \varepsilon_p$ are independent noise (or innovation) variables, $\mathrm{pa}(j)$ denotes the set of parents of node j in the underlying DAG D^0, and $f_j^0(\cdot)$ are real-valued (measurable) functions. We emphasize the true underlying quantities with a superscript "0". We assume in this section that the DAG D^0, or at least a (super-) DAG D^0_{super} which contains D^0 (see Section 4.2.3), is known. As mentioned earlier, our goal is to give a representation of the expected value of the intervention distribution $\mathbb{E}[Y|\mathrm{do}(X = x)]$ for two variables $Y, X \in \{X_1, \ldots, X_p\}$. That is, we want to study the total effect that an intervention at X has on a target variable Y. Let S be a set of variables satisfying the backdoor criterion relative to (X, Y), implying that

$$p(y|\mathrm{do}(X = x)) = \int p(y|X = x, X_S) dP(X_S),$$

where $p(\cdot)$ and $P(\cdot)$ are generic notations for the density or distribution (see Section 4.1.1). Assuming that we can interchange the order of integration (cf. part 6 of Assumption 2) we obtain

$$\mathbb{E}[Y|\mathrm{do}(X = x)] = \int \mathbb{E}[Y|X = x, X_S] dP(X_S). \tag{4.6}$$

This is a function depending on the one-dimensional variable x only and therefore, intuitively, its estimation should not be much exposed to the curse of dimensionality. We will argue below that this is indeed the case.

4.2.1 Marginal integration

Marginal integration is an estimation method which has been primarily designed for additive and structured regression fitting (Linton and Nielsen, 1995). Without any modifications though, it is also suitable for the estimation of $\mathbb{E}[Y|\mathrm{do}(X = x)]$ in (4.6).

Let S be a set of variables satisfying the backdoor criterion relative to (X, Y) (see Section 4.1.1) and denote by s the cardinality of S. We use a nonparametric partial local estimator of the multivariate regression function $m(x, x_S) = \mathbb{E}[Y|X = x, X_S = x_S]$ of the form

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \alpha - \beta(X^{(i)} - x))^2 K_{h_1}(X^{(i)} - x) L_{h_2}(X_S^{(i)} - x_S), \tag{4.7}$$

where $\hat{\alpha} = \hat{\alpha}(x, x_S), \hat{\beta} = \hat{\beta}(x, x_S)$, K and L are two kernel functions and h_1, h_2 the respective bandwidths, that is,

$$K_{h_1}(t) = \frac{1}{h_1} K\left(\frac{t}{h_1}\right), \qquad L_{h_2}(t) = \frac{1}{h_2^s} L\left(\frac{t}{h_2}\right).$$

We obtain the partial local linear estimator at (x, x_S) as $\hat{m}(x, x_S) = \hat{\alpha}(x, x_S)$. We then *integrate* over the variables X_S with the empirical mean and obtain:

$$\hat{\mathbb{E}}[Y|\mathrm{do}(X = x)] \quad = \quad n^{-1} \sum_{k=1}^{n} \hat{m}(x, X_S^{(k)}) \tag{4.8}$$

This is a locally weighted average, with localization through the one-dimensional variable x. For our main theoretical result to hold, we make the following assumptions:

Assumption 2.

1. *The variables X_S have a bounded support $supp(X_S)$.*
2. *The regression function $m(u, u_S) = \mathbb{E}[Y|X = u, X_S = u_S]$ exists and has bounded partial derivatives up to order 2 with respect to u and up to order d with respect to u_S for u in a neighborhood of x and $u_S \in supp(X_S)$.*
3. *The variables X, X_S have a density $p(.,.)$ with respect to Lebesgue measure and $p(u, u_S)$ has bounded partial derivatives up to order 2*

with respect to u and up to order d with respect to u_S. In addition, it holds that

$$\inf_{\substack{u \in x \pm \delta \\ x_S \in supp(X_S)}} p(u, x_S) > 0 \quad \text{for some } \delta > 0.$$

4. *The kernel functions K and L are symmetric with bounded supports and L is an order-d kernel.*

5. *For $\varepsilon = Y - \mathbb{E}[Y|X, X_S]$, it holds that $\mathbb{E}[\varepsilon^4]$ is finite and that $\mathbb{E}[\varepsilon^2|X = x, X_S = x_S]$ is continuous. Furthermore, for a $\delta > 0$, $\mathbb{E}[|\varepsilon|^{2+\delta} \mid X = u]$ is bounded for u in a neighborhood of x.*

6. *There exists $c < \infty$ such that $\mathbb{E}[|Y||X = x, X_S = x_S] \leq c$ for all x_S.*

Note that part 6 of Assumption 2 is only needed for interchanging the order of integration in (4.6). Due to the bounded support of the variables X_S it is not overly restrictive.

As a consequence, the following result from Fan et al. (1998) establishes a convergence rate for the estimator as for one-dimensional nonparametric function estimation.

Theorem 12. *Suppose that Assumption 2 holds for a set S satisfying the backdoor criterion relative to (X, Y) in the DAG D^0 from model* (4.5)*. Consider the estimator in* (4.8)*. Assume that the bandwidths are chosen such that $h_1, h_2 \to 0$ with $nh_1h_2^{2s}/\log^2(n) \to \infty$, $h_2^d/h_1^2 \to 0$, and in addition satisfying $nh_1h_2^s/\log(n) \to \infty$ and $h_1^4 \log(n)/h_2^s \to 0$ (and all these conditions hold when choosing the bandwidths in a properly chosen optimal range, see below). Then,*

$$\hat{\mathbb{E}}[Y|\mathrm{do}(X = x)] - \mathbb{E}[Y|\mathrm{do}(X = x)] = O(h_1^2) + O_P(1/\sqrt{nh_1}).$$

Proof. The statement directly follows from Theorem 1 and Remark 3 in Fan et al. (1998). □

When assuming the smoothness condition $d > s$ for $m(u, u_S)$ with respect to the variable u_S, and when choosing $h_1 \asymp n^{-1/5}$ and $h_2 \asymp n^{-\alpha}$ with $2/(5d) < \alpha < 2/(5s)$ (which requires $d > s$), all the conditions for the bandwidths are satisfied: Theorem 12 then establishes the convergence rate $O(n^{-2/5})$ which matches the optimal rate for estimation of one-dimensional

smooth functions having second derivatives, and such a smoothness condition is assumed for $m(u, u_S)$ with respect to the variable u in part 2 of Assumption 2. Thus, the implication is the important robustness fact that for *any* potentially nonlinear structural equation model satisfying the regularity conditions in Theorem 12, we can estimate the expected value of the intervention distribution with the same accuracy as in nonparametric estimation of a smooth function with one-dimensional argument. We note, as mentioned already in Section 4.1.2, that it would be essentially impossible to estimate the functions f_j in (4.1) in full generality: interestingly, when focusing on inferring the total effect $\mathbb{E}[Y|\mathrm{do}(X = x)]$, the problem is much better posed as demonstrated with our concrete *S-mint* procedure. Furthermore, with the (valid) choice $S = \mathrm{pa}(j_X)$ or an (estimated) superset thereof, one obtains a procedure that is only based on local information in the graph: this turns out to be advantageous, see also Section 4.1.2, particularly when the underlying DAG structure is not correctly specified (see Section 4.5.3). We will report about the performance of such an *S-mint* estimation method in Sections 4.4 and 4.5. Note that the rate of Theorem 12 remains valid (for a slightly modified estimator) if we allow for discrete variables in the parental set of X (Fan et al., 1998).

It is worthwhile to point out that *S-mint* becomes more challenging for inferring multiple variable interventions such as $\mathbb{E}[Y|\mathrm{do}(X_1 = x_1,\ X_2 = x_2)]$: the convergence rate is then of the order $n^{-1/3}$ for a twice differentiable regression function.

Remark 11. *Theorem 12 generalizes to real-valued transformations $t(\cdot)$ of Y. By using the argument as in* (4.6) *and replacing part 6 of Assumption 2 by the corresponding statement for $t(Y)$, we obtain*

$$\begin{aligned}\mathbb{E}[t(Y)|\mathrm{do}(X = x)] &= \int t(y)p(y|\mathrm{do}(X = x))dy \\ &= \int \mathbb{E}[t(Y)|X = x, X_S]dP(X_S).\end{aligned}$$

For example, for $t(y) = y^2$ we obtain second moments and we can estimate the variance $\mathrm{Var}(Y|\mathrm{do}(X = x)) = \mathbb{E}[Y^2|\mathrm{do}(X = x)] - (\mathbb{E}[Y|\mathrm{do}(X = x)])^2$. *Or with the indicator function $t(y) = I(y \le c)$ $(c \in \mathbb{R})$ we obtain a procedure for estimating $\mathbb{P}[Y \le c|\mathrm{do}(X = x)]$ with the same convergence rate as for one-dimensional nonparametric function estimation using marginal integration of $t(Y)$ versus X, X_S.*

Binary treatment and connection to double robustness

For the special but important case with binary treatment $X \in \{0,1\}$ and $X_S \in \mathbb{R}^s$ continuous, we can use marginal integration as well. We can estimate the regression function $m(x, x_S)$ for $x \in \{0,1\}$ by using a kernel estimator based on data with the observed $X^{(k)} = 0$ and $X^{(k)} = 1$, respectively, denoted by $\hat{m}(x, x_S)$ $(x \in \{0,1\})$. We then integrate over x_S with the empirical mean $n^{-1} \sum_{k=1}^{n} \hat{m}(x, X_S^{(k)})$ $(x \in \{0,1\})$. When choosing the bandwidth h_2 (for smoothing over the X_S variables) smaller than for the non-integrated quantity $m(x, x_S)$, and assuming smoothness conditions, we anticipate the $n^{-1/2}$ convergence rate for estimating $\mathbb{E}[Y|\mathrm{do}(X = x)]$ with $x \in \{0,1\}$; see for example Hall and Marron (1987) in the context of nonparametric squared density estimation. We note that this establishes only the optimal parametric convergence rate but does not generally lead to asymptotic efficiency. For the case of binary treatment, semiparametric minimax rates have been established in Robins et al. (2009) and asymptotically efficient methods can be constructed using higher order influence functions (Li et al., 2011) or targeted maximum likelihood estimation (van der Laan and Rose, 2011) which both might be more suitable than marginal integration.

Theorem 12 establishes that *S-mint* is "fully robust" against model misspecification for inferring $\mathbb{E}[Y|\mathrm{do}(X = x)]$ or related quantities as mentioned in Remark 11. The existing framework of double robustness is related to the issue of misspecification and we clarify here the connection. One specifies regression models for $\mathbb{E}[Y|X, X_S] = m(X, X_S)$ for both $X = 0$ and $X = 1$ and a propensity score (Rosenbaum and Rubin, 1983) or inverse probability weighting model (IPW; Robins et al. (1994)): for a binary intervention variable where X encodes "exposure" $(X = 1)$ and "control" $(X = 0)$, the latter is a (often parametric logistic) model for $\mathbb{P}[X = 1|X_S]$. A double robust (DR) estimator for $\mathbb{E}[Y|\mathrm{do}(X = x)]$ requires that either the regression model or the propensity score model is correctly specified. If both of them are misspecified, the DR estimator is inconsistent. Double robustness of the augmented IPW approach has been proved by Scharfstein et al. (1999) and double robustness in general was further developed by many others, see, for example, Bang and Robins (2005). The targeted maximum likelihood estimation (TMLE) framework (van der Laan and Rose, 2011) is also double robust. It uses a second step where the initial estimate is modified in order to make it less biased for the target parameter (e.g., the average causal effect between "exposure" and

"control"). If both, the initial estimator and the treatment mechanism, are consistently estimated, TMLE can be shown to be asymptotically efficient. TMLE with a super-learner or also the approach of higher order influence functions (Li et al., 2011) can deal with a nonparametric model. Robins et al. (2009) prove that $s = \dim(X_S) \leq 2(\beta_{\text{regr}} + \beta_{\text{propens}})$, where $\beta_{\text{'name'}}$ denotes the smoothness of the regression or propensity score function, is a necessary condition for an estimator to achieve the $1/\sqrt{n}$ convergence rate.

Our *S-mint* procedure is related to these nonparametric approaches: it differs though in that it deals with a continuous treatment variable. Similar to the smoothness requirement above we have discussed after Theorem 12 that we can achieve the $n^{-2/5}$ nonparametric optimal rate (when assuming bounded derivatives up to order 2 of the regression function with respect to the treatment variable) if $s = \dim(X_S) < d$, where d plays the role of β_{regr}. The condition $s < d$ is stronger than for the optimal $1/\sqrt{n}$ convergence rate with binary treatment: however, this could be relaxed to the regime $s < 2d$ when invoking Remark 1 in Fan et al. (1998). Therefore, rate optimal estimation with continuous treatment can be achieved under a "comparable" smoothness assumption as in the binary treatment case.

4.2.2 Implementation of marginal integration

Theorem 12 justifies marginal integration as in (4.8) asymptotically. One issue is the choice of the two bandwidths h_1 and h_2: we cannot rely on cross-validation because $\mathbb{E}[Y|\text{do}(X = x)]$ is not a regression function and is not linked to the prediction of a new observation Y_{new}, nor can we use penalized likelihood techniques with, e.g., BIC since $\mathbb{E}[Y|\text{do}(X = x)]$ does not appear in the likelihood. Besides the difficulty of choosing the smoothing parameters, we think that addressing such a smoothing problem will become easier, at least in practice, using an iterative boosting approach (cf. Bühlmann and Yu, 2003; Friedman, 2001).

We propose here a scheme, without complicated tuning of parameters, which we found to be most stable and accurate in extensive simulations. The idea is to elaborate on the estimation of the function $m(x, x_S) = \mathbb{E}[Y|X = x, X_S = x_S]$, from a simple starting point to more complex estimates, while the integration over the variables X_S is done with the empirical mean as in (4.8).

We start with the following simple but useful result.

Proposition 2. *If* $\mathrm{pa}(j_X) = \emptyset$ *or if there are no backdoor paths from* j_X *to* j_Y *in the true DAG* D^0 *from model (4.5), then*

$$\mathbb{E}[Y|\mathrm{do}(X = x)] = \mathbb{E}[Y|X = x].$$

Proof. If there are no backdoor paths from j_X to j_Y, the empty set $S = \emptyset$ satisfies the backdoor criterion relative to (X, Y). The statement then directly follows from the backdoor adjustment formula (4.2). □

We learn from Proposition 2 that in simple situations, a standard one-dimensional regression estimator for $\mathbb{E}[Y|X = x]$ would suffice. On the other hand, we know from the backdoor adjustment formula in (4.6), that we should adjust with the variables X_S. Therefore, it seems natural to use an *additive* regression approximation for $m(x, x_S)$ as a simple starting point. If the assumptions of Proposition 2 hold, such an additive model fit would yield a consistent estimate for the component of the variable x: in fact, it is asymptotically as efficient as when using one-dimensional function estimation for $\mathbb{E}[Y|X = x]$ (Horowitz et al., 2006). If the assumptions of Proposition 2 would not hold, we can still view an additive model fit $\hat{m}_{\mathrm{add}}(x, x_S) = \hat{\mu} + \hat{m}_{\mathrm{add},j_X}(x) + \sum_{j \in S} \hat{m}_{\mathrm{add},j}(x_j)$ as one of the simplest starting points to approximate the more complex function $m(x, x_S)$. When integrating out with the empirical mean as in (4.8), we obtain the estimate $\hat{\mathbb{E}}_{\mathrm{add}}[Y|\mathrm{do}(X = x)] = \hat{\mu} + \hat{m}_{\mathrm{add},j_X}(x)$. As motivated above and backed up by simulations, $\hat{\mu} + \hat{m}_{\mathrm{add},j_X}(x)$ is quite often already a reasonable estimator for $\mathbb{E}[Y|\mathrm{do}(X = x)]$.

In the presence of strong interactions between the variables, the additive approximation may drastically fail though. Thus, we want to implement marginal integration as follows: starting from the additive model fit $\hat{m}_{\mathrm{add}}$, we implement L_2-Boosting with a nonparametric kernel estimator similar to the one in (4.7). More precisely, we compute residuals

$$R_1^{(i)} = Y^{(i)} - \hat{m}_{\mathrm{add}}(X^{(i)}, X_S^{(i)}), \; i = 1, \ldots, n,$$

which, for simplicity, are fitted with a locally constant estimator of the form

$$\hat{\alpha}(x, x_S) = \operatorname*{argmin}_{\alpha} \sum_{i=1}^{n} (R_1^{(i)} - \alpha)^2 K_{h_1}(X^{(i)} - x) L_{h_2}(X_S^{(i)} - x_S). \quad (4.9)$$

The resulting fit is denoted by $\hat{g}_{R_1}(x, x_S) := \hat{\alpha}(x, x_S)$. We add this new function fit to the previous one and compute again residuals, and we then

iterate the procedure b_{stop} times. To summarize, for $b = 1, \ldots, b_{\text{stop}} - 1$,

$$\begin{aligned}
&\hat{m}_1(x, x_S) = \hat{m}_{\text{add}}(x, x_S), \\
&\hat{m}_{b+1}(x, x_S) = \hat{m}_b(x, x_S) + \hat{g}_{R_b}(x, x_S), \\
&R_{b+1}^{(i)} = Y^{(i)} - \hat{m}_{b+1}(X^{(i)}, X_S^{(i)}), \quad i = 1, ..., n.
\end{aligned}$$

The final estimate for the total causal effect is obtained by marginally integrating over the variables X_S with the empirical mean as in (4.8):

$$\hat{\mathbb{E}}[Y|\text{do}(X = x)] = n^{-1} \sum_{k=1}^{n} \hat{m}_{b_{\text{stop}}}(x, X_S^{(k)}).$$

The concrete implementation of the additive model fitting is according to the default from the `R`-package `mgcv`, using penalized thin plate splines and choosing the regularization parameter in the penalty by generalized cross-validation, see, for example, Wood (2003, 2006). The basis dimension for each smooth is set to 10. For the product kernel in (4.9), we choose K to be a Gaussian kernel and L to be a product of Gaussian kernels. The bandwidths h_1 and h_2 in the kernel estimator should be chosen "large" to yield an estimator with low variance but typically high bias. The iterations then reduce the bias. Once we have fixed h_1 and h_2 (and this choice is not very important as long as the bandwidths are "large"), the only regularization parameter is b_{stop}. It is chosen by the following considerations: for each iteration we approximate the sum of the differences to the previous approximation on the set of intervention values $\mathcal{I}$ (typically the nine deciles, see Section 4.5), that is

$$\sum_{x \in \mathcal{I}} |n^{-1} \sum_{k=1}^{n} \hat{g}_{R_b}(x, X_S^{(k)})|. \tag{4.10}$$

When it becomes reasonably "small", and this needs to be specified depending on the context, we stop the boosting procedure. Such an iterative boosting scheme has the advantage that it is more insensitive to the choice of b_{stop} than the original estimator in (4.8) to the specification of the tuning parameters, and in addition, boosting adapts to some extent to different smoothness in different directions (variables). All these ideas are presented at various places in the boosting literature, particularly in Bühlmann and Hothorn (2007), Bühlmann and Yu (2003), and Friedman (2001). In Section 4.4.2 we provide an example of a DAG with backdoor paths, where the additive approximation is incorrect and several boosting iterations are

needed to account for interaction effects between the variables. The implementation of our method is summarized in Algorithm 4: we call it also *S-mint*, and we use it for all our empirical results in Sections 4.4–4.6.

Algorithm 4 S-mint

1: **if** $S = \emptyset$ is a valid adjustment set (for example, if $\mathrm{pa}(j_X) = \emptyset$) **then**
2: Fit an additive regression of Y versus X to obtain $\hat{m}_{\mathrm{add}}$
3: **return** $\hat{m}_{\mathrm{add}}$
4: **else**
5: Fit an additive regression of Y versus X and the adjustment set variables X_S to obtain $\hat{m}_1 = \hat{m}_{\mathrm{add}}$
6: **for** $b = 1, ..., b_{\mathrm{stop}} - 1$ **do**
7: Apply L_2-boosting to capture deviations from an additive regression model:
8: (i) Compute residuals $R_b = Y - \hat{m}_b$
9: (ii) Fit residuals with the kernel estimator (4.9) to obtain $\hat{g}_{R_b}$
10: (iii) Set $\hat{m}_{b+1} = \hat{m}_b + \hat{g}_{R_b}$
11: **end for**
12: **return** Do marginal integration: output

$$\frac{1}{n}\sum_{k=1}^{n} \hat{m}_{b_{\mathrm{stop}}}(x, X_S^{(k)})$$

13: **end if**

We note the following about L_2-boosting: if the initial estimator is a weighted mean $\hat{m}_1(x, x_S) = \sum_{i=1}^n w_i^{(1)}(x, x_S) Y_i$ with $\sum_{i=1}^n w_i^{(1)}(x, x_S) = 1$ (e.g., many additive function estimators are of this form), then, since the kernel estimator $\hat{g}_{R_b}$ in the boosting step 9 is a weighted mean too, $\hat{m}_b(x, x_S) = \sum_{i=1}^n w_i^{(b)}(x, x_S) Y_i$ is a weighted mean. Thus, L_2-boosting has the form of a weighted mean estimator. When using kernel estimation for $\hat{g}_{R_b}$, the boosting estimator $\hat{m}_{b_{\mathrm{stop}}}$ is related to an estimator with a higher order kernel (Di Marzio and Taylor, 2008) which depends on the bandwidth in $\hat{g}_{R_b}$ and the number of boosting iterations in a rather non-explicit way. Establishing the theoretical properties of the L_2-boosting estimator $\hat{\mathbb{E}}[Y|\mathrm{do}(X = x)] = n^{-1}\sum_{k=1}^n \hat{m}_{b_{\mathrm{stop}}}(x, X_S^{(k)})$ is beyond the scope of this work.

4.2.3 Knowledge of a superset of the DAG

It is known that a superset of the parental set $\mathrm{pa}(j_X)$ suffices for the backdoor adjustment in (4.3). To be precise, let

$$S(j_X) \supseteq \mathrm{pa}(j_X) \text{ with } S(j_X) \cap \mathrm{de}(j_X) = \emptyset, \tag{4.11}$$

where $\mathrm{de}(j_X)$ are the descendants of j_X (in the true DAG D^0). For example, $S(j_X)$ could be the parents of X in a superset of the true underlying DAG (a DAG with additional edges relative to the true DAG). We can then choose the adjustment set S in (4.8) as $S(j_X)$ and Theorem 12 still holds true, assuming that the cardinality $|S(j_X)| \leq M < \infty$ is bounded. Thus, with the choice $S = S(j_X)$, we can use marginal integration by marginalizing over the variables $X_{S(j_X)}$.

A prime example where we are provided with a superset $S(j_X) \supseteq \mathrm{pa}(j_X)$ with $S(j_X) \cap \mathrm{de}(j_X) = \emptyset$ is when we know the order of the variables and can deduct an approximate superset of the parents from that. When the variables are ordered with $X_j \prec X_k$ for $j < k$, we would use

$$S(j_X) = \{k; j_X - p_{\max} \leq k < j_X\} \supseteq \mathrm{pa}(j_X), \tag{4.12}$$

where "$\prec$" and p_{max} denote the order relation among the variables and an upper bound on the size of the superset to ensure that $S(j_X) \supseteq \mathrm{pa}(j_X)$.

Corollary 3. *Consider the estimator in* (4.8) *and assume the conditions of Theorem 12 for the variables* Y, X *and* $X_{S(j_X)}$ *with* $S(j_X)$ *in* (4.11) *or* $S(j_X)$ *as in* (4.12) *for ordered variables. Then,*

$$\hat{\mathbb{E}}[Y|\mathrm{do}(X = x)] - \mathbb{E}[Y|\mathrm{do}(X = x)] = O(h_1^2) + O_P(1/\sqrt{nh_1}).$$

Proof. The statement is an immediate consequence of Theorem 12, as $S(j_X)$ in (4.11) and (4.12) satisfy the backdoor criterion relative to (X, Y). □

4.3 Path-based methods

We assume in the following until Section 4.3.5 that we know the true DAG and all true functions and error distributions in the general SEM (4.1). Thus, in contrast to Section 4.2, we have here also knowledge of the entire structure in form of the DAG D^0 (and not only a valid adjustment

set S assumed for Theorem 12). This allows us to infer $\mathbb{E}[Y|\mathrm{do}(X = x)]$ in different ways than the generic *S-mint* regression from Section 4.2. The motivation to look at other methods is driven by potential gains in statistical accuracy when including the additional information of the functional form or of the entire DAG in the structural equation model. We will empirically analyze this issue in Section 4.5.

4.3.1 Entire path-based method from root nodes

Based on the true DAG, the variables can always be ordered such that

$$X_{j_1} \prec X_{j_2} \prec \ldots \prec X_{j_p}.$$

Denote by j_X and j_Y the indices of the variables X and Y, respectively.

If X is not an ancestor of Y, we know that $\mathbb{E}[Y|\mathrm{do}(X = x)] = \mathbb{E}[Y]$. If X is an ancestor of Y it must hold that $j_X < j_Y$. We can then generate the intervention distribution of the random variables

$$X_{j_1} \prec X_{j_2} \prec \ldots \prec Y \,|\, \mathrm{do}(X = x)$$

in the model (4.1) as follows (Pearl, 2009, Definition 3.2.1):

Step 1 Generate $\varepsilon_{j_1}, \ldots, \varepsilon_{j_Y}$.

Step 2 Based on Step 1, recursively generate:

$$\begin{aligned}
&X_{j_1} \leftarrow \varepsilon_{j_1},\\
&X_{j_2} \leftarrow f^0_{j_2}(X_{\mathrm{pa}(j_2)}, \varepsilon_{j_2}),\\
&\ldots,\\
&X_{j_X} \leftarrow x,\\
&\ldots,\\
&X_{j_Y} \leftarrow f^0_{j_Y}(X_{\mathrm{pa}(j_Y)}, \varepsilon_{j_Y}).
\end{aligned}$$

Instead of an analytic expression for $p(Y|\mathrm{do}(X = x))$ by integrating out over the other variables $\{X_{j_k};\ j_k \notin \{j_X, j_Y\}\}$ we rather rely on simulation. We draw B samples $Y^{(1)} = X^{(1)}_{j_Y}, \ldots, Y^{(B)} = X^{(B)}_{j_Y}$ by B independent simulations of Steps 1-2 above and we then approximate, for B large,

$$\mathbb{E}[Y|\mathrm{do}(X = x)] \approx B^{-1} \sum_{b=1}^{B} Y^{(b)}.$$

Furthermore, the simulation technique allows to obtain the distribution of $p(Y|\text{do}(X = x))$ via, for example, density estimation or histogram approximation based on $Y^{(1)}, \ldots, Y^{(B)}$.

The method has an implementation in Algorithm 5 which uses propagation of simulated random variables along directed paths in the DAG. It exploits the entire paths in the DAG from the root nodes to node j_Y corresponding to the random variable Y, see Figure 4.1 for an illustration.

Algorithm 5 Entire path-based algorithm for simulating the intervention distribution

1: If there is no directed path from j_X to j_Y, the interventional and observational quantities coincide: $p(Y|\text{do}(X = x)) \equiv p(Y)$ and $\mathbb{E}[Y|\text{do}(X = x)] \equiv \mathbb{E}[Y]$.
2: If there is a directed path from j_X to j_Y, proceed with steps 3-9.
3: Set $X = X_{j_X} = x$ and delete all in-going arcs into X.
4: Find all directed paths from root nodes (including j_X) to j_Y, and denote them by $p_1, \ldots, p_q$.
5: **for** $b = 1, \ldots, B$ **do**
6: for every path, recursively simulate the corresponding random variables according to the order of the variables in the DAG:

(i) Simulate the random variables corresponding to the root nodes of $p_1, \ldots, p_q$;

(ii) Simulate in each path $p_1, \ldots, p_q$ the random variables following the root nodes; proceed recursively, according to the order of the variables in the DAG.

(iii) Continue with the recursive simulation of random variables until Y is simulated.

7: Store the simulated variable $Y^{(b)}$.
8: **end for**
9: Use the simulated sample $Y^{(1)}, \ldots, Y^{(B)}$ to approximate the intervention distribution $p(y|\text{do}(X = x))$ or its expectation $\mathbb{E}[Y|\text{do}(X = x)]$.

When having estimates of the true DAG, all true functions and error distributions in the additive structural equation model (4.14), we would use the procedure above based on these estimated quantities; for the error distributions, we either use the estimated variances in Gaussian distributions,

or we rely on bootstrapping residuals from the structural equation model (typically with residuals centered around zero).

4.3.2 Partially path-based method with short-cuts

Mainly motivated by computational considerations (see also Section 4.3.3), a modification of the procedure in Algorithm 5 is valid. Instead of considering all paths from root nodes to j_Y (corresponding to variable Y), we only consider all paths from j_X (corresponding to variable X) to j_Y and simulate the random variables on these paths $p'_1, \ldots, p'_m$. Obviously, in comparison to Algorithm 5, $m \leq q$ and every p'_k corresponds to a path p_r for an $r \in \{1, \ldots, q\}$. Every path p'_k is of the form

$$j_X = j_{k,1} \to j_{k,2} \to \ldots \to j_{k,\ell_k - 1} \to j_{k,\ell_k} = j_Y,$$

having length ℓ_k. For recursively simulating the random variables on the paths $p'_1, \ldots, p'_m$ we start with setting

$$X = X_{j_X} \leftarrow x.$$

Then we recursively simulate the random variables corresponding to all the paths $p'_1, \ldots, p'_m$ according to the order of the variables in the DAG. For each of these random variables X_j with $j \in \{p'_1, \ldots, p'_m\}$ and $j \neq j_X$, we need the corresponding parental variables and error terms in

$$X_j \leftarrow f_j^0(X_{\mathrm{pa}(j)}, \varepsilon_j),$$

where for every $k \in \mathrm{pa}(j)$ we set

$$X_k = \begin{cases} \text{the previously simulated value,} & \text{if } k \in \{p'_1, \ldots, p'_m\}, \\ \text{bootstrap resampled } X_k^*, & \text{otherwise,} \end{cases} \tag{4.13}$$

where the bootstrap resampling is with replacement from the entire data. The errors are simulated according to the error distribution. We summarize the procedure in Algorithm 6 and Figure 4.1 provides an illustration

Proposition 3. *Consider the population case in which the bootstrap resampling in* (4.13) *yields the correct distribution of the random variables* $X_1, \ldots, X_p$. *Then, as* $B \to \infty$, *the partially path-based Algorithm 6 yields the correct intervention distribution* $p(y|\mathrm{do}(X = x))$ *and its expected value* $\mathbb{E}[Y|\mathrm{do}(X = x)]$.

Algorithm 6 Partially path-based algorithm for simulating the intervention distribution

1: If there is no directed path from j_X to j_Y, the interventional and observational quantities coincide: $p(Y|\text{do}(X = x)) \equiv p(Y)$ and $\mathbb{E}[Y|\text{do}(X = x)] \equiv \mathbb{E}[Y]$.
2: If there is a directed path from j_X to j_Y, proceed with steps 3-9.
3: Set $X = X_{j_X} = x$ and delete all in-going arcs into X.
4: Find all directed paths from j_X to j_Y and denote them by $p'_1, ..., p'_m$.
5: **for** $b = 1, \ldots, B$ **do**
6: for every path, recursively simulate the corresponding random variables according to the order of the variables in the DAG:

(i) Simulate in each path $p'_1, \ldots, p'_m$ the random variables following the node j_X; proceed recursively as described in (4.13) according to the order of the variables in the DAG.

(ii) Continue with the recursive simulation of random variables until Y is simulated.

7: Store the simulated variable $Y^{(b)}$.
8: **end for**
9: Use the simulated sample $Y^{(1)}, \ldots, Y^{(B)}$ to approximate the intervention distribution $p(y|\text{do}(X = x))$ or its expectation $\mathbb{E}[Y|\text{do}(X = x)]$.

Proof. The statement of Proposition 3 directly follows from the definition of the intervention distribution in a structural equation model. □

The same comment as in Section 4.3.1 applies here: when having estimates of the quantities in the additive structural equation model (4.14), we would use Algorithm 6 based on the plugged-in estimates. The computational benefit of using Algorithm 6 instead of Algorithm 5 is illustrated in Figure 4.8.

4.3.3 Degree of localness

We can classify the different methods according to the degree of which the entire or only a small (local) fraction of the DAG is used. Algorithm 5 is a rather global procedure as it uses entire paths from root nodes to j_Y. Only when j_Y is close to the relevant root nodes, the method does involve

a smaller aspect of the DAG. Algorithm 6 is of semi-local nature as it does not require to consider paths going from root nodes to j_Y: it only considers paths from j_X to j_Y and all parental variables along these paths. The *S-mint* method based on marginal integration described in Section 4.2 and Theorem 12 is of very local character as it only requires the knowledge of Y, X and the parental set $\mathrm{pa}(j_X)$ (or a superset thereof) but no further information about paths from j_X to j_Y.

In the presence of estimation errors, a local method might be more "reliable" as only a smaller fraction of the DAG needs to be approximately correct; global methods, in contrast, require that entire paths in the DAG are approximately correct. The local versus global issue is illustrated qualitatively in Figure 4.1, and empirical results about statistical accuracy of the various methods are given in Section 4.5.

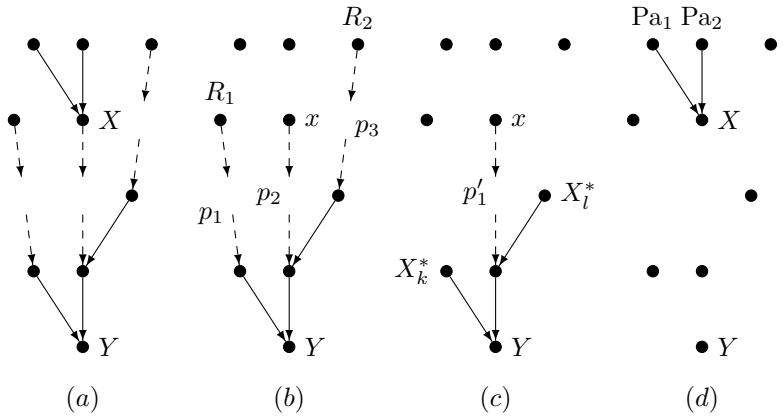


Figure 4.1: (a) True DAG D^0. (b) Illustration of Algorithm 1. X is set to x, the roots R_1, R_2 and all paths from the root nodes to Y are enumerated (here: p_1, p_2, p_3). The interventional distribution at node Y is obtained by propagating samples along the three paths. (c) Illustration of Algorithm 2. X is set to x and all directed paths from X to Y are labeled (here: p_1'). In order to obtain the interventional distribution at node Y, samples are propagated along the path p_1' and bootstrap resampled X_k^* and X_l^* are used according to (4.13). (d) Illustration of the *S-mint* method with adjustment set $S = \mathrm{pa}(j_X)$: it only uses information about Y, X and the parents of X (here: $\mathrm{Pa}_1, \mathrm{Pa}_2$).

4.3.4 Estimation of DAG, edge functions and error distributions

With observational data, in general, it is impossible to infer the true underlying DAG D^0 in the structural equation model (4.5), or its parental sets, even as sample size tends to infinity. One can only estimate the Markov equivalence class of the true DAG, assuming faithfulness of the data-generating distribution, see, for example, Spirtes et al. (2000), Pearl (2000), Chickering (2002), Kalisch and Bühlmann (2007), van de Geer and Bühlmann (2013), and Bühlmann (2013). The latter three references focus on the high-dimensional Gaussian scenario with the number of random variables $p \gg n$ but assuming a sparsity condition in terms of the maximal degree of the skeleton of the DAG D^0. The edge functions and error variances can then be estimated for every DAG member in the Markov equivalence class by pursuing regression of a variable versus its parents.

However, there are interesting exceptions regarding identifiability of the DAG from the observational distribution. For nonlinear structural equation models with additive error terms, it is possible to infer the true underlying DAG from infinitely many observational data (Hoyer et al., 2009; Peters et al., 2014). Various methods have been proposed to infer the true underlying DAG D^0 and its corresponding functions $f_j^0(\cdot)$ and error distributions of the ε_j's: see, for example, Imoto et al. (2002), Hoyer et al. (2009), Peters et al. (2014), Bühlmann et al. (2014), van de Geer (2014), and Nowzohour and Bühlmann (2016) (the fourth and fifth references are considering high-dimensional scenarios). Another interesting class of models where the DAG D^0 can be identified from the observational data distribution are linear structural equation models with non-Gaussian noise (Shimizu et al., 2006), or with Gaussian noise but equal or approximately equal error variances (Loh and Bühlmann, 2014; Peters and Bühlmann, 2014; van de Geer and Bühlmann, 2013) (the first and third references are considering the high-dimensional setting).

As an example of a model with identifiable structure (DAG D^0) we can specialize (4.5) to an additive structural equation model of the form

$$X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} f_{jk}^0(X_k) + \varepsilon_j, \ j = 1, \ldots, p, \tag{4.14}$$

where $\varepsilon_1, \ldots, \varepsilon_p$ are independent with $\varepsilon_j \sim \mathcal{N}(0, (\sigma_j^0)^2)$, and the true underlying DAG is denoted by D^0. This model is used for all numerical comparisons of the *S-mint* procedure and the path-based algorithms in

Section 4.5. Estimation of the unknown quantities D^0, f^0_{jk} and error variances $(\sigma^0_j)^2$ can be done with the CAM method (see Chapter 2 for details), which will be used for the empirical results in Section 4.5.4 in connection with the two-stage procedure *est S-mint* that will be introduced next.

4.3.5 Two-stage procedure: *est S-mint*

If the order of the variables or (a superset of) the parental set is unknown, we have to estimate it from observational data; this leads to the following two-stage procedure described here for the case where the parental set $\mathrm{pa}(j_X)$ is identifiable:

Stage 1 Estimate a superset of the parental set $S(j_X)$ (defined in (4.11)) from observational data.

Stage 2 Based on the estimate $\hat{S}(j_X)$, run *S-mint* regression with adjustment set $S = \hat{S}(j_X)$.

Even if in Stage 1 one would also obtain estimates of functions in a specified SEM besides an estimate of $S(j_X)$, we would not use the estimated functions in Stage 2. We present empirical results for the *est S-mint* procedure in connection with the CAM method for Stage 1 for estimating a valid adjustment set $S(j_X)$ in Section 4.5.4.

If the parental set $\mathrm{pa}(j_X)$ is not identifiable (see Section 4.3.4), one could apply Stage 1 to obtain a set $\{\hat{S}(j_X)^{(1)}, \ldots, \hat{S}(j_X)^{(c_j)}\}$ such that the parental sets from each of the equivalent DAGs would be contained in at least one of the $\hat{S}(j_X)^{(k)}$ for some k. Stage 2 would then be performed for all estimates $\{\hat{S}(j_X)^{(1)}, \ldots, \hat{S}(j_X)^{(c_j)}\}$ and one could then derive bounds of the quantity $\mathbb{E}[Y|\mathrm{do}(X = x)]$ in the spirit of the approach of Maathuis et al. (2009).

In Section 4.5.5 we will give some intuition why the two stage *est S-mint* is often leading to better and more reliable results than (at least some) other methods which rely on path-based estimation.

4.4 Empirical results: non-additive structural equation models

In this section we provide simple proof-of-concept examples for the generality of the proposed *S-mint* estimation method (Algorithm 4). In particular, the robustness of *S-mint* is experimentally validated for models where the structural equation model is not additive as in (4.14) but given in its general form (4.5). We make a naive comparison to path-based methods which are inconsistent due to incorrect specification of the model in Section 4.4.1. However, taking the view of classical robustness (cf. Hampel et al., 2011), we consider a complementary and interesting issue in Section 4.5: namely the "efficiency" of a robust procedure in comparison to other methods relying on the correct model specification.

In Section 4.4.1 we empirically show that the path-based methods based on the wrong additive model assumption in (4.14) may fail even in the absence of backdoor paths where the *S-mint* method boils down to estimation of an additive model. In Section 4.4.2 we add backdoor paths to the graph and a strong interaction term to the corresponding structural equation model. We then empirically show that *S-mint* manages to approximate the true causal effect, whereas fitting only an additive regression fails. Section 4.4.3 contains an example that demonstrates a good performance of *S-mint* even in the presence of non-additive noise in the structural equation model. Finally, Section 4.4.4 empirically illustrates issues with the fixed choice of the bandwidths in the product kernel in (4.9) in some cases. Unless stated differently, we set both bandwidths to 0.4.

4.4.1 Causal effects in the absence of backdoor paths

First let us illustrate the sensitivity of the path-based methods with respect to model specification, using a simple example of a 4-node graph with no backdoor paths between $X_1 = X$ and Y (see Figure 4.2).

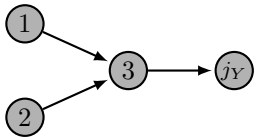


Figure 4.2: Example of a DAG without backdoor paths.

We consider a corresponding (non-additive) structural equation model of the form

$$\begin{aligned} X_1 &\leftarrow \varepsilon_1 \\ X_2 &\leftarrow \varepsilon_2 \\ X_3 &\leftarrow \cos(4 \cdot (X_1 + X_2)) \cdot \exp(X_1/2 + X_2/4) + \varepsilon_3 \\ Y &\leftarrow \cos(X_3) \cdot \exp(X_3/4) + \varepsilon_4, \end{aligned} \tag{4.15}$$

where $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_1 = \sigma_2 = 0.7$ and $\sigma_3 = \sigma_4 = 0.2$. We generate n samples from this model. From Proposition 2 we know that for $j \in \{1, 2, 3\}$, fitting an additive regression of Y versus X_j and $X_{\mathrm{pa}(j)}$ suffices to obtain the causal effect $\mathbb{E}[Y|\mathrm{do}(X_j = x)]$, that is, all causal effects can be readily estimated with an additive model. Our goal is to infer $\mathbb{E}[Y|\mathrm{do}(X_1 = x)]$, based on $n = 500$ and $n = 10'000$ samples of the joint distribution of the 4 nodes. The results are displayed in Figure 4.3. We consider the entire path-based Algorithm 5 (and Algorithm 6 as well, not shown) assuming an additive structural equation model as in (4.14). We impressively see that this approach is exposed to model misspecification while *S-mint* (in this case simply fitting of an additive model, that is, $b_{\mathrm{stop}} = 1$ with the number of additional boosting iterations equaling zero) is not and leads to reliable and correct results. We included two settings; $n = 500$ to be consistent with the settings in the numerical study from Section 4.5 and $n = 10000$ to demonstrate that the failure of the path-based methods is not a small sample size but an inconsistency phenomenon.

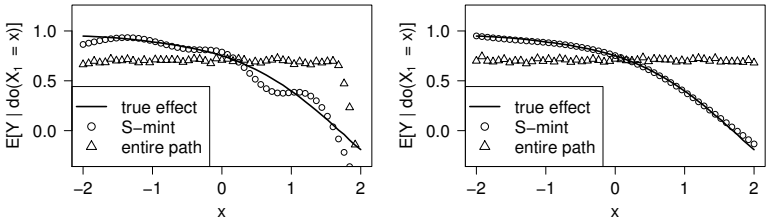


Figure 4.3: *S-mint* regression estimates of $\mathbb{E}[Y|\mathrm{do}(X_1 = x)]$ for the model in (4.15), with $S = S(j_X = 1) = \emptyset$, based on one representative sample each for sample sizes $n = 500$ (left) and $n = 10'000$ (right). *S-mint* regression is consistent while the entire path-based method with a misspecified additive SEM (Algorithm 5) is not. The relative squared errors (over the 51 points x) are 0.013 for *S-mint* regression and 6.239 for the entire path-based method, both for $n = 10000$.

4.4.2 Causal effects in the presence of backdoor paths

We consider a slight (but crucial) modification of the above DAG that has been proposed by Linbo Wang and Mathias Drton through private communication. We consider the 4-node graph from Figure 4.2 with additional edges $X_1 \to Y$ and $X_2 \to Y$ and corresponding structural equation model

$$\begin{aligned} X_1 &\leftarrow \varepsilon_1 \\ X_2 &\leftarrow \varepsilon_2 \\ X_3 &\leftarrow X_1 + X_2 + \varepsilon_3 \\ Y &\leftarrow X_1 \cdot X_2 \cdot X_3 + \varepsilon_4 \end{aligned} \tag{4.16}$$

where $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_1 = \sigma_2 = 0.7$ and $\sigma_3 = \sigma_4 = 0.2$. Note that this modification introduces two backdoor paths from X_3 to Y. The goal is to estimate the causal effect $\mathbb{E}[Y|\mathrm{do}(X_3 = x)]$ using the *S-mint* estimation procedure introduced in Algorithm 4 with different numbers of boosting iterations. In Figure 4.4 one clearly sees that the additive approximation (with no additional boosting iterations) fails to approximate the total causal effect. It is not able to capture the full interaction term $X_1 \cdot X_2 \cdot X_3$. Adding boosting iterations significantly improves the approximation also for the small sample size $n = 500$ (Ernest and Bühlmann, 2015, Figure 3).

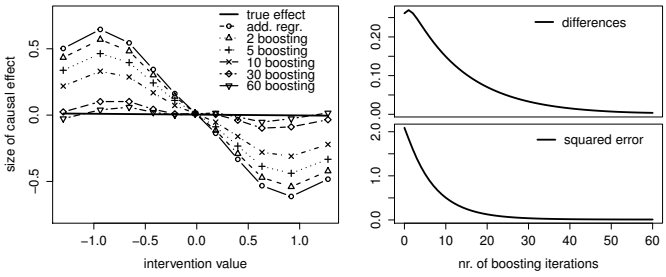


Figure 4.4: Approximation of the causal effect $\mathbb{E}[Y|\mathrm{do}(X_3 = x)]$ in model (4.16) with *S-mint* regression for additive model fit (starting value) and various boosting iterations (left), absolute differences between consecutive boosting iterations as in (4.10) (upper right) and integrated squared error for approximating the true effect as a function of boosting iterations (lower right). The boosting iterations in the *S-mint* procedure account for interactions between the variables. The adjustment set is chosen as the parental set of X_3, that is $S(j_X = 3) = \{1, 2\}$. The results are based on one representative sample of size $n = 10000$.

4.4.3 Causal effects in the presence of non-additive noise

Theorem 12 does not put any explicit restrictions on the noise structure in the structural equation model. In particular, *S-mint* also works well in the case of non-additive noise. As an example, we consider the causal graph and SEM from Section 4.4.2 but replace the structural equation corresponding to Y in (4.16) with

$$Y \leftarrow \exp(X_1) \cdot \cos(X_2 \cdot X_3 + \varepsilon_4). \tag{4.17}$$

The goal is again to estimate the causal effect $\mathbb{E}[Y|\mathrm{do}(X_3 = x)]$ based on $n = 500$ observed samples of the joint distribution. Figure 4.5 shows that *S-mint* yields a close approximation to the true causal effect.

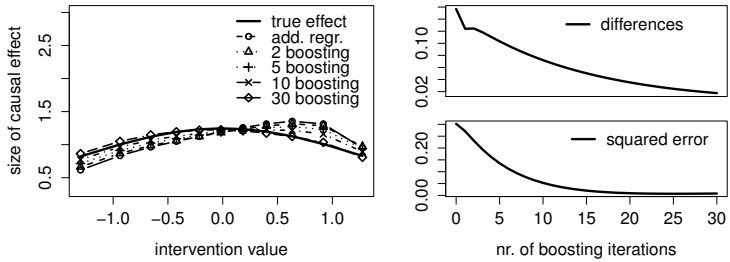


Figure 4.5: Approximation of the causal effect $\mathbb{E}[Y|\mathrm{do}(X_3 = x)]$ in model (4.17) exhibiting non-additive noise in the structural equation model, with *S-mint* regression for additive model fit (starting value) and various boosting iterations (left). Absolute differences between consecutive boosting iterations as in (4.10) (upper right) and integrated squared error for approximating the true effect as a function of boosting iterations (lower right). The adjustment set is chosen as the parental set of X_3, that is $S(j_X = 3) = \{1, 2\}$. The results are based on one representative sample of size $n = 500$.

4.4.4 Choice of the bandwidth

Theorem 12 provides an asymptotic result but does not specify the bandwidths h_1 and h_2 in the finite sample case. In particular, the same fixed choice of h_2 for all variables in the adjustment set S can be suboptimal in some situations. As an example let us consider the graph and structural

equations from Section 4.4.2 where we replace one equation in (4.16) by

$$Y \leftarrow X_1 + \sin(X_2 \cdot X_3) + \varepsilon_4. \tag{4.18}$$

The goal is to estimate the causal effect $\mathbb{E}[Y|\mathrm{do}(X_3 = x)]$ based on $n = 500$ samples of the joint distribution. Inspecting the scatterplots of Y versus X_1, X_2 and X_3 (Figure 4.6) suggests that the bandwidth $h_2^{(1)}$ corresponding to X_1 should be larger than the bandwidth $h_2^{(2)}$ corresponding to X_2.

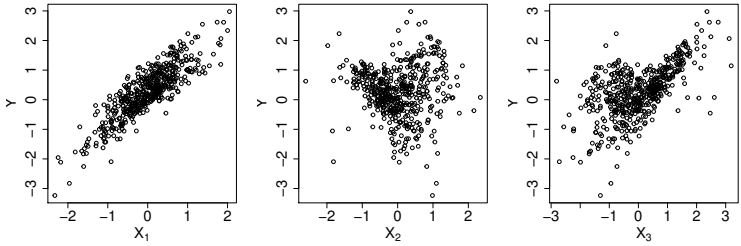


Figure 4.6: Scatterplots of the data from model (4.18) of Y versus X_1, X_2 and X_3. They reveal a difference in wigglyness.

Figure 4.7 depicts the corresponding approximated causal effects using the *S-mint* method for a fixed bandwidth $h_2 = (h_2^{(1)}, h_2^{(2)}) = (0.4, 0.4)$ and for a variable bandwidth $h_2 = (h_2^{(1)}, h_2^{(2)}) = (0.8, 0.4)$, respectively.

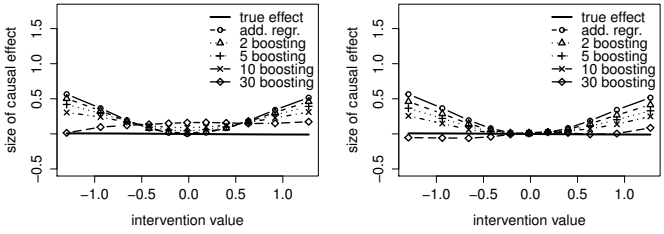


Figure 4.7: Approximation of the causal effect $\mathbb{E}[Y|\mathrm{do}(X_3 = x)]$ in model (4.18). The adjustment set is chosen as the parental set of X_3, that is $S(j_X = 3) = \{1, 2\}$ and with corresponding fixed bandwidths $h_2^{(1)} = h_2^{(2)} = 0.4$ (left) and varying bandwidths $h_2^{(1)} = 0.8$ and $h_2^{(2)} = 0.4$ (right). The results are based on one representative sample of size $n = 500$.

The approximation with the variable bandwidth outperforms the approximation with the fixed bandwidth. Adaptive bandwidths choice methods as proposed by Polzehl and Spokoiny (2000) might be suitable, at the price of a more complicated and hence more variable estimation scheme.

4.5 Empirical results: Additive structural equation models

The goal of the numerical experiments in this section is to quantify the estimation accuracy of the total causal effect $\mathbb{E}[Y|\mathrm{do}(X = x)]$ for two variables $X, Y \in \{X_1, ..., X_p\}$ such that Y is a descendant of X (if Y is an ancestor of X, then the interventional expectation corresponds to the observational expectation $\mathbb{E}[Y]$). We consider in this section only additive structural equation models as in (4.14). This allows for a comparison of the *S-mint* method and the path-based methods.

For the *S-mint* regression, we use the implementation described in Section 4.2.2. The kernel functions K and L in the *S-mint* procedure are chosen to be a Gaussian kernel with bandwidth h_1 and a product of Gaussian kernels with bandwidth h_2, respectively. For simplicity, in the style of Fan et al. (1998), we choose h_1 and h_2 as 0.5 times the empirical standard deviation of the respective covariables in all of our simulations in this section. We use the following two criteria for b_{stop}, that is, as an automated stopping criterion for the boosting iterations:

1. Stop if an iteration changes the approximation by less than 1%. That is, the integrated difference (4.10) is less than 1% of the integrated previous approximation.

2. Stop if the absolute difference between two consecutive integrated differences is less than 5% of the initial integrated difference.

When using the path-based methods from Section 4.3, we estimate the functions f_j^0 by additive functions using the `R`-package `mgcv` with default values (and thus use the knowledge of the form of the nonlinear functions in the SEM).

We test the performance of four different methods: *S-mint* with parental sets (Algorithm 4) with the stopping of boosting iterations as described above, additive regression with parental sets (first step of *S-mint*, without

additional boosting iterations), entire path-based method from root nodes (Algorithm 5) and partially path-based method with short-cuts (Algorithm 6). The reference effect $\mathbb{E}[Y|\mathrm{do}(X = x)]$ is computed using Algorithm 5 with known (true) functions $f^0_{j,k}$ and error variances $(\sigma^0_j)^2$ based on $5n$ samples.

Since in a nonlinear structural equation model (in contrast to a linear structural equation model) $\mathbb{E}[Y|\mathrm{do}(X = x)]$ is a nonlinear function of the intervention value x, we compute the interventional expectation for several values x: typically, for the nine deciles $d_1(X), ..., d_9(X)$ of X. To compare the estimation accuracy of the three methods on DAG D, we compute a relative squared error $e(D)$ over all considered pairs (X, Y) (for details see below), denoted by $\mathcal{L}$, and over all intervention values $d_1(X), ..., d_9(X)$ as

$$e(D) = \frac{\sum_{(X,Y)\in\mathcal{L}} \sum_{i=1}^{9} \left(\hat{\mathbb{E}}[Y|\mathrm{do}(X = d_i(X))] - \mathbb{E}^0[Y|\mathrm{do}(X = d_i(X))]\right)^2}{\sum_{(X,Y)\in\mathcal{L}} \sum_{i=1}^{9} \left(\mathbb{E}^0[Y|\mathrm{do}(X = d_i(X))]\right)^2}. \tag{4.19}$$

Typically, we repeat every experiment on $N = 50$ or $N = 100$ random DAGs (described in Section 4.5.1) and record the relative error $e(D)$ of all methods for each repetition.

4.5.1 Data simulation

To simulate data we first fix a causal order π^0 of the variables, that is $X_{\pi^0(1)} \prec X_{\pi^0(2)} \prec \cdots \prec X_{\pi^0(p)}$ and include each of the $\binom{p}{2}$ possible directed edges, independently of each other, with probability p_c. In the sparse setting we typically choose $p_c = 2/(p-1)$ which yields an expected number of p edges in the resulting DAG. Based on the causal structure of the graph we then build the structural equation model. We simulate from the additive structural equation model (4.14), where every edge $k \to j$ in the DAG is associated with a nonlinear function $f^0_{j,k}$ in the structural equation model. We use two function types:

1. edge functions $f^0_{j,k}$ drawn from a Gaussian process with a Gaussian kernel with bandwidth one
2. sigmoid-type edge functions of the form $f^0_{j,k}(x) = a \cdot \frac{b\cdot(x+c)}{1+|b\cdot(x+c)|}$ with $a \sim \mathrm{Exp}(4) + 1$, $b \sim \mathrm{Unif}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathrm{Unif}([-2, 2])$.

All variables with empty parental set (root nodes in the DAG) follow a Gaussian distribution with mean zero and standard deviation which is uniformly distributed in the interval $[1, \sqrt{2}]$. To all remaining variables we add Gaussian noise with standard deviation uniformly distributed in $[1/5, \sqrt{2}/5]$. Note that both simulation settings correspond to the ones used in Bühlmann et al. (2014), see Section 2.6.

4.5.2 Estimation of causal effects for known graphs

In this section we compare the different methods in terms of estimation accuracy and CPU time consumption for known underlying DAGs D^0. To that end we generate random DAGs with $p = 10$ variables and simulate $n = 500$ samples of the joint distribution applying the simulation procedure introduced in Section 4.5.1. We then select all index pairs (k, j) such that there exists a directed path from X_k to X_j and estimate the causal effect $\mathbb{E}[X_j|\mathrm{do}(X_k)]$ for all k, j on the nine deciles of X_k.

The experiment is done for two different levels of sparsity, a sparse graph with an expected number of p edges and a non-sparse graph with an expected number of $4p$ edges. We record the relative squared error (4.19) and the CPU time consumption, both averaged over all index pairs, for $N = 100$ ($N = 20$ in the dense setting, respectively) different DAGs D^0. The results are displayed in Figure 4.8 for the sigmoid-type edge functions and in Figure 4.9 for the Gaussian process-type edge functions.

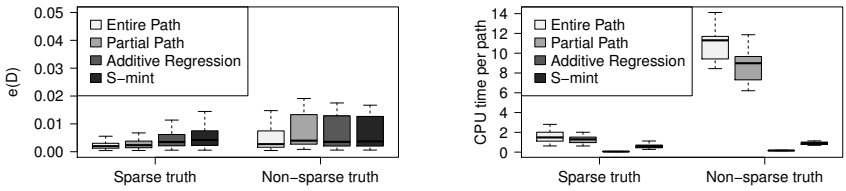


Figure 4.8: Comparison of the performance of the methods in terms of relative squared error as in (4.19) (left) and CPU time consumption (right) for the case where the true DAGs D^0 are known and the edge functions belong to the sigmoid-type setting. The adjustment set is $S = \mathrm{pa}(X_k)$ for additive regression and *S-mint*. Number of variables $p = 10$ and sample size $n = 500$.

The method based on the entire paths (Algorithm 5) yields the smallest errors followed by the path-based methods with short-cuts (Algorithm 6). The *S-mint* and additive regression exhibit a slightly worse performance.

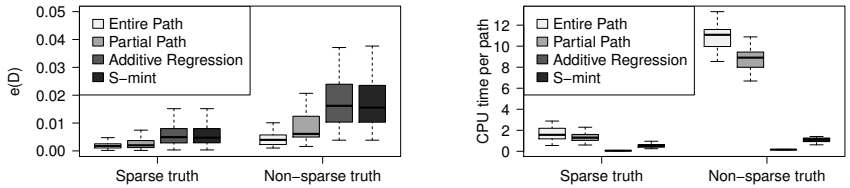


Figure 4.9: Comparison of the performance of the methods in terms of relative squared error as in (4.19) (left) and CPU time consumption (right) for the case where the true DAGs D^0 are known and the edge functions are drawn from a Gaussian process with bandwidth one. The adjustment set is $S = \text{pa}(X_k)$ for additive regression and *S-mint*. Number of variables $p = 10$ and sample size $n = 500$.

This finding can be explained by the fact that the path-based methods benefit from the full (and correct) structural information of the DAG whereas the *S-mint* and additive regression methods only use local information (cf. Section 4.3.3). For the monotone sigmoid-type function class, additive regression provides a very good approximation to the true causal effect even in dense settings. For both settings we observe that the boosting iterations in *S-mint* do not improve the additive approximation substantially.

In terms of CPU time consumption, *S-mint* and additive regression outperform the path-based methods. Additive regression is particularly fast as it only requires the fit of one nonparametric additive regression of X_j versus X_k and $X_{\text{pa}(k)}$ whereas the path-based methods each require one nonparametric additive model fit for every node on all the traversed paths. As the set of paths in the partially path-based method is a subset of the one in the entire path-based method (cf. Section 4.3.2 and Figure 4.1), the partially path-based method needs less model fits which explains the reduction of time consumption. In particular, both *S-mint* and additive regression are computationally feasible for computing $\mathbb{E}[X_j|\text{do}(X_k)]$ for all pairs (k, j), even when p is large and in the thousands assuming that the cardinality of the corresponding adjustment sets is reasonably small.

4.5.3 Estimation of causal effects for perturbed graphs

In the previous section we demonstrated that the two path-based methods exhibit a better performance than *S-mint* and the additive regression approximation if causal effects are estimated based on the underlying true DAG D^0.

We will now focus on the more realistic situation in which we are only provided with a partially correct DAG $\tilde{D}$. We model this by constructing a set of modified DAGs $\{\tilde{D}_{h_r}\}_{r\in\mathcal{K}}$ with pre-specified (fixed) structural Hamming distances $\{h_r\}_{r\in\mathcal{K}}$ to the true DAG D^0, where $\mathcal{K} = \{1, 2, \ldots, 6\}$ and the corresponding $\{h_r\}_{r\in\mathcal{K}}$ are described in Figures 4.10 and 4.11. To do so, we use the following rule: starting from D^0 with $p = 50$ nodes, for each $r \in \mathcal{K}$, we randomly remove and add $\frac{h_r}{2}$ edges each to obtain $\tilde{D}_{h_r}$. The structural Hamming distance between D^0 and the perturbed graph $\tilde{D}_{h_r}$ is then equal to h_r, and an percentage of $1-\frac{h_r}{2|E|}$ edges in $\tilde{D}_{h_r}$ are still correct, where $|E|$ denotes the expected number of edges in the DAG D^0. Note that this modification may change the order of the variables (especially for large values of h_r). We randomly choose $20 = |\mathcal{L}|$ index pairs (k, j) such that there exists a directed path from X_k to X_j in D^0, but now consider the problem of estimating the total causal effect $\mathbb{E}[X_j|\mathrm{do}(X_k)]$ based on the perturbed graph $\tilde{D}_{h_r}$ for the adjustment sets or the paths, respectively (and based on sample size $n = 500$ as in Section 4.5.2). For every $r \in \mathcal{K}$, this is repeated $N = 100$ times and in each repetition, we record the relative squared error $e(D)$ in (4.19). As before, we distinguish between a sparse graph with an expected number of 50 edges and a non-sparse graph with an expected number of 200 edges and we use both simulation settings described in Section 4.5.1 for generating the edge functions f^0. The results are shown in Figures 4.10 and 4.11.

For both, the sparse and non-sparse settings, one observes that the larger the structural Hamming distance (or equivalently, the smaller the percentage of correctly specified edges in D^0), the better is the performance of *S-mint* and additive regression in comparison with the path-based methods. That is, both methods are substantially more robust with respect to possible misspecifications of edges in the graph. This may be explained by the different degrees of localness (cf. Section 4.3.3) of the respective methods. For the two local methods we can hope to have approximately correct information in the parental set of X_k even if the modified DAG is far away from the true DAG D^0 in terms of the structural Hamming distance. For the path-based methods however, randomly removing edges may break one or several of the traversed paths which results in causal information being partially or fully lost. This effect is most evident in the two sparse settings. A similar behavior is also observed in Figure 4.12.

Note that except for the true DAG D^0, the performance of the partially path-based method is at least as good as for the entire path-based method. The shortcut introduced in Algorithm 6 does not only yield computational

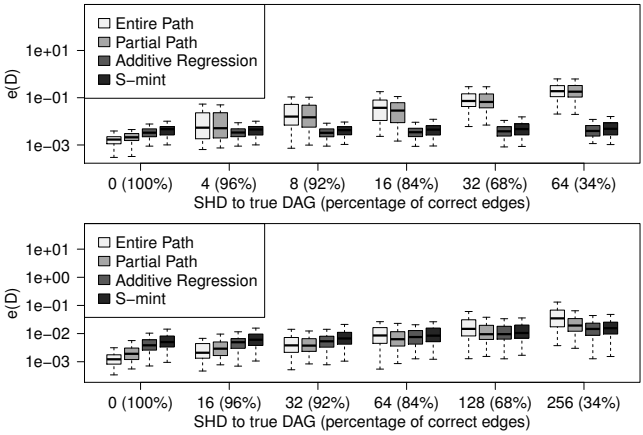


Figure 4.10: The plots compare the relative squared error performance of the three methods on a set of modified DAGs $\{\tilde{D}_{h_r}\}_{r\in\mathcal{K}}$ with given structural Hamming distances $\{h_r\}_{r\in\mathcal{K}}$ to the true DAG D^0 (or equivalently, with a given percentage of correct edges) for the sigmoid-type additive structural equation model. The top and bottom panels show the relative squared error error $e(D)$ (4.19) in a sparse and dense setting, respectively. The larger the structural Hamming distance h_r between the modified DAG $\tilde{D}_{h_r}$ and the true DAG D^0, the better is the performance of *S-mint* with parental sets in comparison with the two path-based methods. Number of variables $p = 50$ and sample size $n = 500$.

savings but also improves (relative to the entire path-based Algorithm 5) statistical estimation accuracy of causal effects in incorrect DAGs. Again, a possible explanation for this observation is that the partially path-based method acts more locally and thus is less affected by edge perturbations.

4.5.4 Estimation of causal effects for estimated graphs

We now turn our attention to the case where the goal is to compute causal effects on a DAG $\hat{D}$ that has been estimated by a structure learning algorithm (while still relying on a correct model specification). In conjunction with *S-mint* regression, this is then the *est S-mint* method described in Section 4.3.5.

We generate $N = 50$ random DAGs with $p = 20$ nodes for different numbers n of observational data, which are simulated according to the procedure in Section 4.5.1.

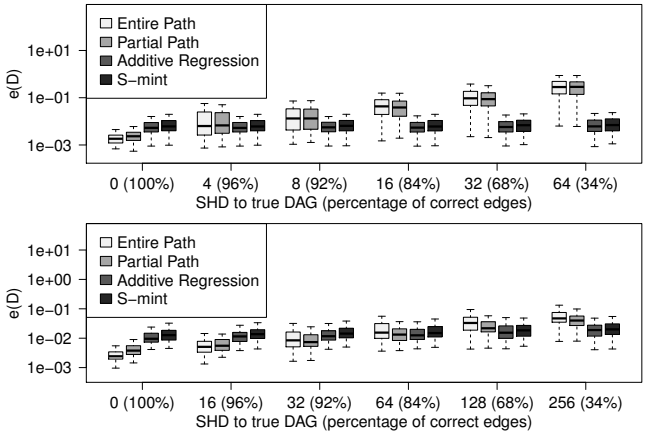


Figure 4.11: The plots compare the relative squared error performance of the three methods on a set of modified DAGs $\{\tilde{D}_{h_r}\}_{r\in\mathcal{K}}$ with given structural Hamming distances $\{h_r\}_{r\in\mathcal{K}}$ to the true DAG D^0 (or equivalently, with a given percentage of correct edges) for the Gaussian process-type additive structural equation models. The top and bottom panels show the relative squared error $e(D)$ (4.19) in a sparse and dense setting, respectively. The larger the structural Hamming distance h_r between the modified DAG $\tilde{D}_{h_r}$ and the true DAG D^0, the better is the performance of *S-mint* with parental sets in comparison with the two path-based methods. Number of variables $p = 50$ and sample size $n = 500$.

Using the knowledge that the structural equation model is additive as in (4.14), we apply the recently proposed CAM method (Bühlmann et al., 2014) for estimation of the true underlying DAG D^0 (which is identifiable from the observational distribution). For details, see Chapter 2. The implementation is according to the `R`-package `CAM`. Regarding the algorithmic details, we use the following in the three steps:

1. Preliminary neighborhood selection to restrict the number of potential parents per node: set to a maximum of 10 by default;

2. Estimation of the correct order by greedy search: we use 6 basis functions per parent to fit the additive model;

3. Optional: Pruning of the DAG by feature selection to keep only the significant edges, where we use the default level $\alpha = 0.001$.

After having estimated a DAG $\hat{D}$ with the above procedure, we randomly

select $10 = |\mathcal{L}|$ index pairs (k, j) such that there exists a directed path from X_k to X_j in the true DAG D^0 and approximate the total causal effect $\mathbb{E}[X_j|\mathrm{do}(X_k)]$ based on the estimated graph $\hat{D}$. Figure 4.12 displays the relative squared errors as defined in (4.19).

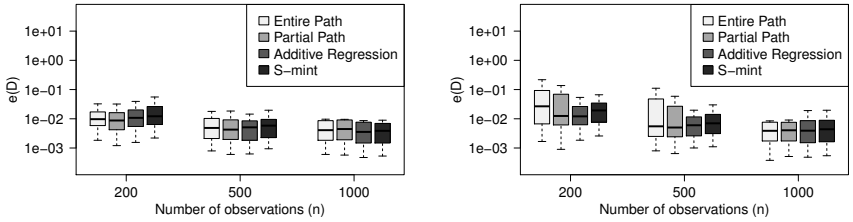


Figure 4.12: Sigmoid-type additive structural equation models. Relative squared error performance as in (4.19), for different numbers of observations (n), computed on graphs that have been estimated using the CAM method described in Section 2.5. The algorithm has been applied without the pruning step (left) and with the pruning step (right). We use the estimated parental sets as adjustment sets and the number of variables is $p = 20$. The *S-mint* regression corresponds to *est S-mint* as described in Section 4.3.5.

All four methods show a similar performance with respect to relative squared error on the DAGs that are obtained applying the CAM method without feature selection. These DAGs mainly represent the causal order of the variables but otherwise are densely connected. An incorrectly specified order of the variables (e.g., for small sample sizes n) seems to comparably affect the *S-mint* and additive regression with parental sets and the path-based methods. If the sample size increases, the estimated graph $\hat{D}$ is closer to the true graph D^0 which improves the estimation accuracy of causal effects for all the four methods.

The two path-based methods approximate the causal effects more accurately on the DAGs that are obtained without feature selection, that is, pruning the DAG is not advantageous for the estimation accuracy of causal effects, at least for a small number of observations. However, the pruning step yields vast computational savings for the two path-based methods as demonstrated in Figure 4.13. The *S-mint* regression is very fast in both settings and pruning the DAG before estimating the causal effects only has a minor effect on the time consumption and estimation accuracy.

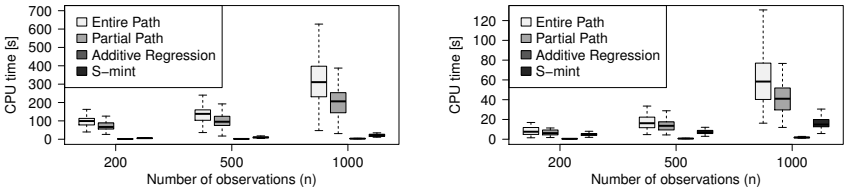


Figure 4.13: Sigmoid-type additive structural equation models. CPU time performance for $n = 500$ for $N = 50$ graphs of $p = 20$ variables that have been estimated using the CAM method described in Section 2.5 with and without pruning step. Pruning the DAG yields vast computational savings for the two path-based methods. *S-mint* and additive regression are barely affected by the pruning step and are considerably faster than the two path-based methods in both scenarios.

4.5.5 Summary of the empirical results, and the advantage of the two-stage *est S-mint* method

With respect to statistical accuracy, measured with the relative squared error as in (4.19), we find that *S-mint* and additive regression are substantially more robust against incorrectness of the true underlying DAG (or against a wrong order of the variables) and against model misspecification, in comparison to the alternative path-based methods. The latter robustness of *S-mint* is rigorously backed-up by our theory in Theorem 12 and Corollary 3 whereas the former seems to be due to the higher degree of localness as described in Section 4.3.3. As a consequence, the proposed two-stage *est S-mint* (Section 4.3.5) where we first estimate the order of the variables or the structure of the DAG (or the Markov equivalence class of DAGs) and subsequently perform *S-mint* is expected in general to lead to reasonably accurate results (which are empirically quantified above for some settings). Only when the DAG is perfectly known and the model correctly specified (here by an additive structural equation model), which is a rather unrealistic assumption for practical applications, the path-based methods were found to have a slight advantage. Thus, we recover here a typical robustness phenomenon against model misspecification of our nonparametric and more "model-free" *S-mint* regression procedure.

Our empirical findings support the use of *est S-mint*, namely the combination of a structured nonparametric (or parametric) approach for estimating the DAG (or its equivalence class) in the first stage and using the robust and fully nonparametric *S-mint* procedure in the second stage. The second

stage leads to a clear gain in robustness whereas the efficiency loss in case of correctly specified models is marginal or even minimal.

Regarding computational efficiency, *S-mint* and in particular also the additive regression approximation are massively faster than the path-based procedures making them feasible for larger scales where the number of variables is in the thousands.

4.6 Real data application

In this section we want to provide two examples for the application of our methodology to real data. We use gene expression data from the isoprenoid biosynthesis in *Arabidopsis thaliana* (Wille et al., 2004). The data consists of $n = 118$ gene expression measurements from $p = 39$ genes. In the original work the authors try to infer connections between the individual genes in the network using Gaussian graphical modeling. Our goal is to find the strongest causal connections between the individual genes. We do not standardize the original data but adjust the bandwidths in *S-mint* by scaling with the standard deviations of the corresponding variables.

4.6.1 Estimation and error control for causal connections between and within the pathways

We first turn our attention to the whole isoprenoid biosynthesis dataset and want to find the causal effects within and between the different pathways, with an error control for false positive selections. To be able to compute the causal effects we have to estimate a causal network. In order to do that we use the CAM method described in Section 2.5.

We estimate a DAG using CAM with the default settings. We then apply the *S-mint* procedure with parental sets obtained from the estimated DAG (which corresponds to the *est S-mint* procedure from Section 4.3.5) to rank the total causal effects according to their strength. We define the relative causal strength $\mathrm{CS}^{\mathrm{rel}}_{k\to j}$ of an intervention $X_j|\mathrm{do}(X_k)$ as a sum of relative distances of observational and interventional expectation for different intervention values divided by the range of the intervention values, that is,

$$\mathrm{CS}^{\mathrm{rel}}_{k\to j} = \frac{1}{R_k(d)} \sum_{i=1}^{9} \frac{|\mathbb{E}[X_j] - \mathbb{E}[X_j|\mathrm{do}(X_k = d_i)]|}{|\mathbb{E}[X_j]|},$$

where we choose $d_1(X_k), ..., d_9(X_k)$ to be the nine deciles of X_k and we denote their range by $R_k(d) = d_9(X_k) - d_1(X_k)$.

To control the number of false positives (i.e., falsely selected strong causal effects) we use stability selection (Meinshausen and Bühlmann, 2010), which provides (conservative) error control under a so-called (and uncheckable) exchangeability condition. We randomly select 100 subsamples of size $n/2 = 59$ and repeat the procedure above 100 times. For each run, we record the indices of the top 30 ranked causal strengths. At the end we keep all index pairs that have been selected at least 66 times in the 100 runs as this leads to an expected number of falsely selected edges (false positives) which is less or equal to 2 (Meinshausen and Bühlmann, 2010). The graphical representation of the network in Figure 4.14 is based on Wille et al. (2004). The dotted arcs represent the underlying metabolic network (known from biology), the six red solid arcs correspond to the stable index pairs found by *est S-mint* with stability selection.

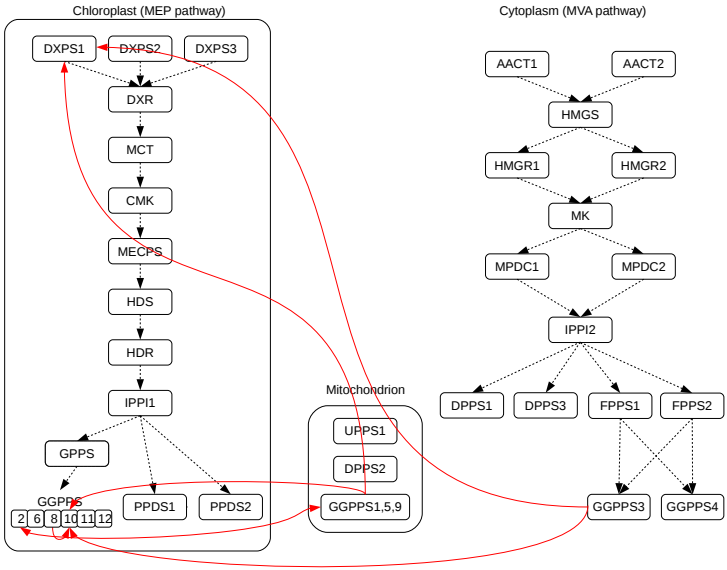


Figure 4.14: Stable edges (with stability selection) for the *Arabidopsis thaliana* dataset. The dotted arcs represent the metabolic network, the red solid arcs the stable total causal effects found by the *est S-mint* method.

None of the stable edges are opposite to the direction of the metabolic network. In particular, we found strong total causal effects between the GGPPS variables in the MEP pathway, MVA pathway and mitochondrion. Note that in this section we heavily rely on model assumption (4.14) as the CAM method for estimating a DAG assumes additivity of the parents. Therefore we cannot fully exploit the advantage of the *S-mint* method that it works for arbitrary non-additive models (4.5) (but we would hope to be somewhat less sensitive to model misspecification than with path-based methods, see for example Figures 4.10 and 4.11).

4.6.2 Estimation and error control of strong causal connections within the MEP pathway

We now want to present a possible way of exploiting the very general model assumptions of *S-mint*. If the underlying order and an approximate graph structure are known *a priori*, we can use this information to proceed with *S-mint* using the order information as described in Corollary 3. This relieves us from any model assumptions on the functional connections between two variables (e.g., linearity, additivity, etc.).

To give an example, let us focus on the genes in the MEP pathway (black box in Figure 4.14). The goal is to find the strongest total causal effects within this pathway. The metabolic network (dotted arcs) is providing us with an order of the variables which we use for *S-mint* regression as follows: we choose the adjustment set $S(j_X)$ in (4.12) by going three levels back ($p_{\max} = 3$) in the causal order (to achieve a reasonably sized set), for example, the adjustment set for CMK is DXPS1, DXPS2, DXPS3, DXR, MCT, whereas the adjustment set for GPPS is HDS, HDR, IPPI1. We cannot use the full set of all ancestors as there are only $n/2 = 59$ data points to fit the nonparametric additive regression and marginal integration, as we again use stability selection based on subsampling for controlling false positive selections as described in the previous section. For each among the 100 subsampling runs we record the top 10 ranked index pairs and keep the ones that are selected at least 65 times out of 100 repetitions. This results in an expected number of false positives being less than 1 (Meinshausen and Bühlmann, 2010). The stable edges are shown in Figure 4.15. One of the four edges corresponds to an edge in the metabolic pathway. We find that the upper part of the pathway contains the strongest total causal effects and therefore may be an interesting target for intervention experiments.

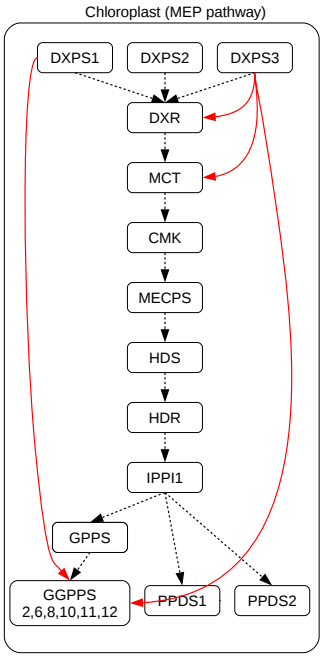


Figure 4.15: Stable edges (with stability selection) for the MEP pathway in the *Arabidopsis thaliana* dataset. The dotted arcs represent the metabolic network whereas the red solid arcs denote the top ranked causal effects found by *S-mint* with adjustment sets chosen from the order of the metabolic network structure by considering all ancestors up to three levels back.

4.7 Conclusions

We considered the problem of estimating expected values of intervention distributions, also known as total causal effects, from observational data. A first main result (Theorem 12 and Corollary 3) says that if we know the local parental variables or a superset thereof (e.g., from the order of the variables), there is no need to base estimation and computations on a causal graph. In fact, we can directly infer the expected values of single-intervention distributions via marginal integration: we call the procedure *S-mint*. This result holds for any nonlinear and non-additive structural

equation model apart from mild smoothness and regularity conditions. Hence, from another point of view, *S-mint* estimation of expected values of single intervention distributions is a fully nonparametric technique and thus robust against model misspecification of the functional form of the structural equations. We propose an L_2-boosting approach for *S-mint* which is easy to use without complicated tuning of parameters and yields good empirical results.

We complement the robustness view-point by empirical results indicating that *S-mint* also works reasonably well when the DAG- or order-structure is misspecified to a certain extent, as it will be the case when we estimate these quantities from data; in fact, *S-mint* regression is substantially more robust than methods which follow all directed paths in the DAG to infer causal effects. This suggests that the two-stage *est S-mint* procedure is most reliable for causal inference from observational data: first estimate the DAG- or order-structure (or equivalence classes thereof) and second, subsequently pursue *S-mint* regression. In addition, such a procedure is computationally much faster than methods which exploit directed paths in (estimated) DAGs.

Chapter 5

Conclusion and Outlook

In the previous chapters, we have addressed the problems of identifiability, structure learning and estimation of causal effects under interventions for selected classes of semiparametric and nonparametric structural equation models. In this concluding chapter we briefly recapitulate the most important contributions and suggest routes for further research.

Structure learning for causal additive models

In Chapter 2, we developed a structure learning methodology for the class of causal additive models (CAMs) based on (restricted) maximum likelihood estimation. The methodology was shown to be consistent in low- and high-dimensional settings and allows for misspecification of the error distributions. In addition, we have proposed the first algorithm to learn low- and high-dimensional CAMs from observational data and have provided an efficient implementation in the `R`-package `CAM` that can deal with up to thousands of variables. With that, we have made an important contribution to structure learning for an interesting class of structural equation models.

In comparison with methods that assume an underlying linear model, our proposed methodology demonstrated a superior performance on data that was generated from a CAM (cf. Section 2.6.3), and a quite comparable performance on data generated from a linear Gaussian SEM (cf. Section 2.6.5). Yet, further work is needed to validate the CAM methodology

on real data with known ground truth, and to compare its performance to state-of-the-art structure learning methods such as the PC-algorithm or GES.

Practically, interesting algorithmic extensions of the CAM methodology could be to incorporate background knowledge on specific causal relations (e.g., from intervention experiments or expert knowledge), or to assess the significance of individual estimated edges. The former extension is relatively straightforward. One could simply add the edge orientations imposed by the background knowledge before the *IncEdge* step of CAM. The latter would be particularly relevant in cases where the model is misspecified (e.g., in the case of linear functions in the model) and the associated DAG is only identifiable up to an equivalence class.

Identifiability and estimation of partially linear models

A major limitation of both, linear SEMs and CAMs is that they rely on the assumption of exclusivity of the functions. In Chapter 3, we addressed this limitation by considering the more general class of partially linear additive SEMs with Gaussian noise (PLSEMs). We presented a graphical, a transformational, a functional and a causal ordering characterization of the identifiability of PLSEMs. The former two are generalizations of well-known graphical and transformational characterization results for Markov equivalence classes. The latter two precisely specify how single nonlinear additive functions impose restrictions on the set of potential underlying causal structures. Based on our theoretical results, we developed an efficient score-based methodology that, for a given PLSEM, lists all equivalent PLSEMs. The presented results are the first that systematically address the identifiability and estimation of (partially linear) additive models with Gaussian noise and non-exclusive functional type.

Our characterizations of PLSEMs and the proposed estimation methodology constitute an interesting first step towards the development of a structure learning methodology for the class of PLSEMs. Naively, one could use the CAM algorithm to obtain an initial DAG estimate and then search for equivalent PLSEMs with the methodology proposed in Section 3.3. This may be a reasonable approach if most of the functions in the PLSEM are nonlinear, but is generally not expected to yield accurate results, especially in the presence of many linear functions. As an alternative for the general setting, one could implement a heuristic greedy search strategy that visits a number of randomly chosen distributionally equivalent DAGs in each

step of the greedy search (cf. Castelo and Kocka, 2003).

Future research in the area of identifiability of PLSEMs could pursue a further reduction of the number of assumptions. For example, by considering partially linear additive SEMs with arbitrary error distributions or partially linear SEMs with a potentially non-additive nonparametric part. For the former, it would be crucial to understand the interplay of non-exclusive functional types and arbitrary noise distributions. This may allow us to explicitly characterize cases with special identifiability properties as done in Zhang and Hyvärinen (2009) for the bivariate case. For the latter, it would be interesting to study whether a non-additive nonparametric part imposes similar restrictions on the underlying DAG of a PLSEM as an additive nonparametric part. Similar characterizations of identifiability might be derived for these more general partially linear SEMs. The modification of our proofs towards the proposed generalizations, however, is non-trivial as we strongly rely on both, the Gaussianity of the noise and the additivity of the functions.

Estimation of total causal effects in nonparametric models

In Chapter 4, we examined what is possible in general nonparametric structural equation models where one is not willing to make any kind of structural assumptions on the functions or distributional assumptions on the noise. While structure learning for this general class of SEMs suffers from the curse of dimensionality, we showed that under suitable assumptions and (approximate) knowledge of the causal structure, total causal effects under single variable interventions can be predicted without entering the curse of dimensionality. In fact, a specific marginal integration estimator achieves the optimal univariate rate of convergence for nonparametric function estimation. We proposed a reasonably robust implementation of our methodology based on an additive approximation and L_2-boosting without complicated tuning of parameters.

Our contribution has an important conceptual implication: instead of trying to learn a causal graph under relatively strict model assumptions, it sometimes could be more promising (and substantially easier) to rely on knowledge of an approximate causal structure (e.g., based on expert knowledge) to estimate total causal effects. An interesting area where such an approximate knowledge of the causal structure can naturally be assumed is in a time series context. The extension of Theorem 12 and Algorithm 4 to the setting of time series is addressed in Li et al. (2016).

Future work could focus on the (experimental) validation of the proposed methodology on real datasets. Thereby, it would be relevant to assess the performance of the methodology in the prediction of strong (top-ranked) causal effects, for example, in the spirit of Maathuis et al. (2010). Also, it would be interesting to examine the gain of using a nonparametric estimation technique over one that relies on parametric assumptions such as linearity of the structural equations.

List of figures

Introduction

Structure learning for CAMs

Identifiability & estimation of PLSEMs

Estimation of causal effects in nonparametric SEMs

List of tables

Identifiability & estimation of PLSEMs

List of algorithms

Identifiability & estimation of PLSEMs

Estimation of causal effects in nonparametric SEMs

Bibliography

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). „A characterization of Markov equivalence classes for acyclic digraphs“. *Annals of Statistics* 25 (2), pp. 505–541.

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), pp. 444–455.

Bang, H. and Robins, J. (2005). „Doubly robust estimation in missing data and causal inference models“. *Biometrics* 61, pp. 962–972.

Bollen, K. A. (1998). *Structural equation models*. Wiley Online Library.

Breiman, L. and Friedman, J. H. (1985). „Estimating optimal transformations for multiple regression and correlation“. *Journal of the American Statistical Association* 80 (391), pp. 580–598.

Bühlmann, P. (2013). „Causal statistical inference in high dimensions“. *Mathematical Methods of Operations Research* 77, pp. 357–370.

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

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

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

Bühlmann, P., Peters, J., and Ernest, J. (2014). „CAM: Causal Additive Models, high-dimensional order search and penalized regression“. *Annals of Statistics* 42 (6), pp. 2526–2556. DOI: 10.1214/14-AOS1260.

Castelo, R. and Kocka, T. (2003). „On Inclusion-driven Learning of Bayesian Networks“. *Journal of Machine Learning Research* 4, pp. 527–574.

Chickering, D. M. (1995). „A Transformational Characterization of Equivalent Bayesian Network Structures“. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann, pp. 87–98.

– (2002). „Optimal structure identification with greedy search“. *Journal of Machine Learning Research* 3, pp. 507–554.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. (2012). „Learning high-dimensional directed acyclic graphs with latent and selection variables“. *Annals of Statistics* 40, pp. 294–321.

Dawid, A. P. (2000). „Causal inference without counterfactuals“. *Journal of the American Statistical Association* 95 (450), pp. 407–424.

Di Marzio, M. and Taylor, C. (2008). „On boosting kernel regression“. *Journal of Statistical Planning and Inference* 138, pp. 2483–2498.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). „Assumptions of IV Methods for Observational Epidemiology“. *Statistical Science* 25 (1), pp. 22–40.

Editorial (2010). „Cause and effect“. *Nature Methods* 7, p. 243.

Ernest, J., Rothenhäusler, D., and Bühlmann, P. (2016). *Causal inference in partially linear structural equation models: identifiability and estimation*. arXiv:1607.05980.

Ernest, J. and Bühlmann, P. (2015). „Marginal integration for nonparametric causal inference“. *Electronic Journal of Statistics* 9 (2), pp. 3155–3194. DOI: 10.1214/15-EJS1075.

Fan, J., Härdle, W., Mammen, E., et al. (1998). „Direct estimation of low-dimensional components in additive models“. *Annals of Statistics* 26 (3), pp. 943–971.

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

Friedman, N. (2004). „Inferring cellular networks using probabilistic graphical models“. *Science* 303 (5659), pp. 799–805.

Glass, T. A., Goodman, S. N., Hernán, M. A., and Samet, J. M. (2013). „Causal Inference in Public Health“. *Annual Review of Public Health* 34 (1), pp. 61–75.

Greenland, S., Pearl, J., and Robins, J. M. (1999). „Causal diagrams for epidemiologic research“. *Epidemiology* 10 (1), pp. 37–48.

Hall, P. and Marron, J. (1987). „Estimation of integrated squared density derivatives“. *Statistics & Probability Letters* 6, pp. 109–115.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Second. New York: Springer.

Hauser, A. and Bühlmann, P. (2012). „Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs“. *The Journal of Machine Learning Research* 13 (1), pp. 2409–2464.

– (2014). „Two optimal strategies for active learning of causal models from interventional data“. *International Journal of Approximate Reasoning* 55, pp. 926–939.

– (2015). „Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs“. *Journal of the Royal Statistical Society, Series B* 77, pp. 291–318.

He, Y.-B. and Geng., Z. (2008). „Active learning of causal networks with intervention experiments and optimal designs“. *Journal of Machine Learning Research* 9, pp. 2523–2547.

Horowitz, J., Klemelä, J., and Mammen, E. (2006). „Optimal estimation in additive regression models“. *Bernoulli* 12, pp. 271–298.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). „Model-based boosting 2.0.“ *Journal of Machine Learning Research* 11, pp. 2109–2113.

Hoyer, P. O., Hyvarinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., and Shimizu, S. (2008). „Causal discovery of linear acyclic models

with arbitrary distributions". In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 282–289.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). „Nonlinear causal discovery with additive noise models". In *Advances in Neural Information Processing Systems 21 (NIPS)*. Red Hook, NY: Curran, pp. 689–696.

Husmeier, D. (2003). „Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks". *Bioinformatics* 19, pp. 2271–2282.

Hyttinen, A., Hoyer, P., Eberhardt, F., and Jarvisalo, M. (2013). „Discovering Cyclic Causal Models with Latent Variables: A General SAT-Based Procedure". In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 301–310.

Hyttinen, A., Eberhardt, F., and Jarvisalo, M. (2014). „Constraint-based Causal Discovery: Conflict Resolution with Answer Set Programming". In *Proceedings of the 30th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 340–349.

Imoto, S., Goto, T., and Miyano, S. (2002). „Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression". In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*. Vol. 7, pp. 175–186.

Janzing, D., Peters, J., Mooij, J. M., and Schölkopf, B. (2009). „Identifying confounders using additive noise models". In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 249–257.

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

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), pp. 1–26.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. (2008). „Discovering Cyclic Causal Models by Independent Components Analysis“. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 366–374.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

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

Li, L., Tchetgen, E. T., Vaart, A. van der, and Robins, J. (2011). „Higher order inference on a treatment effect under low regularity conditions“. *Statistics & Probability Letters* 81, pp. 821–828.

Li, S., Ernest, J., and Bühlmann, P. (2016). *Nonparametric causal inference from observational time series through marginal integration*. arXiv: 1606.04431.

Linton, O. and Nielsen, J. P. (1995). „A kernel method of estimating structured nonparametric regression based on marginal integration“. *Biometrika*, pp. 93–100.

Loh, P. and Bühlmann, P. (2014). „High-dimensional learning of linear causal networks via inverse covariance estimation“. *Journal of Machine Learning Research* 15 (1), pp. 3065–3105.

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

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

Maathuis, M. H. and Colombo, D. (2015). „A generalized back-door criterion“. *The Annals of Statistics* 43 (3), pp. 1060–1088.

Mammen, E. and Park, B. U. (2006). „A simple smooth backfitting method for additive models“. *Annals of Statistics* 34, pp. 2252–2271.

Marra, G. and Wood, S. (2011). „Practical variable selection for generalized additive models“. *Computational Statistics & Data Analysis* 55, pp. 2372–2387.

Meek, C. (1995). „Causal Inference and Causal Explanation with Background Knowledge“. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann, pp. 403–410.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). „High-Dimensional Additive Modeling“. *Annals of Statistics* 37, pp. 3779–3821.

Meinshausen, N. and Bühlmann, P. (2006). „High-dimensional graphs and variable selection with the Lasso“. *Annals of Statistics* 34, pp. 1436–1462.

– (2010). „Stability Selection (with discussion)“. *Journal of the Royal Statistical Society, Series B* 72, pp. 417–473.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. (2009). „Regression by Dependence Minimization and its Application to Causal Inference“. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 745–752.

Mooij, J., Janzing, D., Heskes, T., and Schölkopf, B. (2011). „On Causal Discovery with Cyclic Additive Noise Models“. In *Advances in Neural Information Processing Systems 24 (NIPS)*. Red Hook, NY: Curran, pp. 639–647.

Mooij, J. and Heskes, T. (2013). „Cyclic Causal Discovery from Continuous Equilibrium Data“. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 431–439.

Nandy, P., Hauser, A., and Maathuis, M. H. (2016). *High-dimensional consistency in score-based and hybrid structure learning*. arXiv:1507.02608.

Nowzohour, C. and Bühlmann, P. (2016). „Score-based causal learning in additive noise models“. *Statistics* 50 (3), pp. 471–485.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

– (2009). *Causality: Models, Reasoning and Inference*. 2nd edition. Cambridge University Press.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2016). *Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs*. arXiv:1606.06903.

Peters, J. and Bühlmann, P. (2014). „Identifiability of Gaussian Structural Equation Models with Equal Error Variances“. *Biometrika* 101, pp. 219–228.

– (2015). „Structural Intervention Distance (SID) for Evaluating Causal Graphs“. *Neural Computation* 27 (3), pp. 771–799.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2014). „Causal Discovery with Continuous Additive Noise Models“. *Journal of Machine Learning Research* 15, pp. 2009–2053.

Peters, J., Bühlmann, P., and Meinshausen, N. (2015). „Causal inference using invariant prediction: identification and confidence intervals“. *Journal of the Royal Statistical Society, Series B*.

Polzehl, J. and Spokoiny, V. (2000). „Adaptive weights smoothing with applications to image restoration“. *Journal of the Royal Statistical Society, Series B* 62, pp. 335–354.

Ramsey, J., Zhang, J., and Spirtes, P. (2006). „Adjacency-Faithfulness and Conservative Causal Inference“. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. Arlington, VA: AUAI Press, pp. 401–408.

Ramsey, J., Hanson, S., Hanson, C., Halchenko, Y., Poldrack, R., and Glymour, C. (2010). „Six problems for causal inference from fMRI“. *NeuroImage* 49 (2), pp. 1545 –1558.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). „Sparse additive models“. *Journal of the Royal Statistical Society, Series B* 71, pp. 1009–1030.

Rényi, A. (1959). „On measures of dependence“. *Acta Mathematica Hungarica* 10, pp. 441–451.

Richardson, T. (1996). „A discovery algorithm for directed cyclic graphs“. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann, pp. 454–461.

Richardson, T. and Spirtes, P. (2002). „Ancestral graph Markov models“. *The Annals of Statistics* 30 (4), pp. 962–1030.

Robins, J., Tchetgen, E. T., Li, L., and Vaart, A. van der (2009). „Semiparametric minimax rates“. *Electronic Journal of Statistics* 3, pp. 1305–1321.

Robins, J., Rotnitzky, A., and Zhao, L. (1994). „Estimation of regression coefficients when some of the regressors are not always observed“. *Journal of the American Statistical Association* 89, pp. 846–866.

Rosenbaum, P. and Rubin, D. (1983). „The central role of the propensity score in observational studies for causal effects“. *Biometrika* 70, pp. 41–55.

Rothenhäusler, D., Heinze, C., Peters, J., and Meinshausen, N. (2015). „BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions“. In *Advances in Neural Information Processing Systems 28*. Red Hook, NY: Curran, pp. 1513–1521.

Rubin, D. (2005). „Causal inference using potential outcomes“. *Journal of the American Statistical Association* 100 (469).

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). „Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion)“. *Journal of the American Statistical Association* 94, pp. 1096–1146.

Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007). „Learning graphical model structure using L1-regularization paths“. In *Proceedings of the National Conference on Artificial Intelligence*. Vol. 22. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 1278.

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

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). „DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model“. *Journal of Machine Learning Research* 12, pp. 1225–1248.

Shojaie, A. and Michailidis, G. (2010). „Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs“. *Biometrika* 97, pp. 519–538.

Shpitser, I., Richardson, T. S., and Robins, J. M. (2011). „An efficient algorithm for computing interventional distributions in latent variable causal models“. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 661–670.

Smith, V. A., Jarvis, E. D., and Hartemink, A. J. (2002). „Evaluating functional network inference using simulations of complex biological systems“. *Bioinformatics* 18 (suppl 1), S216–S224.

Song, L., Fukumizu, K., and Gretton, A. (2013). „Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models“. *Signal Processing Magazine, IEEE* 30, pp. 98–111.

Spirtes, P. (1995). „Directed cyclic graphical representations of feedback models“. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann, pp. 491–499.

– (2010). „Introduction to causal inference“. *The Journal of Machine Learning Research* 11, pp. 1643–1662.

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

Spirtes, P. and Glymour, C. (1991). „An Algorithm for Fast Recovery of Sparse Causal Graphs“. 9 (1), pp. 62–72.

Spirtes, P. and Zhang, K. (2016). „Causal discovery and inference: concepts and recent methodological advances“. *Applied Informatics* 3 (1), pp. 1–28.

Statnikov, A., Henaff, M., Lytkin, N. I., and Aliferis, C. F. (2012). „New methods for separating causes from effects in genomics data“. *BMC genomics* 13 (Suppl 8), S22.

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

Teyssier, M. and Koller, D. (2005). „Ordering-based search: a simple and effective algorithm for learning Bayesian networks“. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*. Arlington, VA: AUAI Press, pp. 584–590.

Tibshirani, R. (1996). „Regression Shrinkage and Selection via the Lasso“. *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288.

Triantafillou, S. and Tsamardinos, I. (2015). „Constraint-based Causal Discovery from Multiple Interventions over Overlapping Variable Sets“. *Journal of Machine Learning Research* 16, pp. 2147–2205.

Triantafilou, S., Tsamardinos, I., and Tollis, I. G. (2010). „Learning Causal Structure from Overlapping Variable Sets“. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 9, pp. 860–867.

van de Geer, S. (2014). „On the uniform convergence of empirical norms and inner products, with application to causal inference“. *Electronic Journal of Statistics* 8 (1), pp. 543–574.

van de Geer, S. and Bühlmann, P. (2013). „ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs“. *Annals of Statistics* 41 (2), pp. 536–567.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). „The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso)“. *Electronic Journal of Statistics* 5, pp. 688–749.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning. Causal Inference for Observational and Experimental Data*. New York: Springer.

Verma, T. and Pearl, J. (1990). „Equivalence and Synthesis of Causal Models“. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (UAI)*. Corvallis, OR: AUAI Press, pp. 220–227.

Voorman, A., Shojaie, A., and Witten, D. (2014). „Graph estimation with joint additive models“. *Biometrika* 101 (1), pp. 85–101.

Wainwright, M. (2009). „Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)“. *IEEE Transactions on Information Theory* 55, pp. 2183–2202.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., Rohr, P. von, Thiele, L., et al. (2004). „Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana“. *Genome Biology* 5 (11), R92.

Wood, S. N. (2003). „Thin-plate regression splines“. *Journal of the Royal Statistical Society (B)* 65 (1), pp. 95–114.

– (2006). *Generalized Additive Models: An Introduction with R.* CRC.

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A., and Jarvis, E. (2004). „Advances to Bayesian network inference for generating causal networks from observational biological data". *Bioinformatics* 20, pp. 3594–3603.

Yuan, M. and Lin, Y. (2006). „Model selection and estimation in regression with grouped variables". *Journal of the Royal Statistical Society, Series B* 69, pp. 49–67.

Zhang, C.-H. and Huang, J. (2008). „The sparsity and bias of the Lasso selection in high-dimensional linear regression". *Annals of Statistics* 36, pp. 1567–1594.

Zhang, J. (2008). „On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias". *Artificial Intelligence* 172 (16), pp. 1873–1896.

Zhang, J. and Spirtes, P. (2008). „Detection of Unfaithfulness and Robust Causal Inference". *Minds and Machines* 18 (2), pp. 239–271.

Zhang, K. and Hyvärinen, A. (2009). „On the Identifiability of the Post-Nonlinear Causal Model". In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI).* Corvallis, OR: AUAI Press, pp. 647–655.

Zhao, P. and Yu, B. (2006). „On model selection consistency of Lasso". *Journal of Machine Learning Research* 7, pp. 2541–2563.

Zou, H. (2006). „The adaptive Lasso and its oracle properties". *Journal of the American Statistical Association* 101, pp. 1418–1429.