Foundations of structural causal models with cycles and latent variables

Bongers, S. ; Forré, P. ; Peters, J ; Mooij, J.M.

DOI
10.1214/21-AOS2064

Publication date
2021

Document Version
Final published version

Published in
The Annals of Statistics

The Annals of Statistics
2021, Vol. 49, No. 5, 2885–2915
https://doi.org/10.1214/21-AOS2064

FOUNDATIONS OF STRUCTURAL CAUSAL MODELS WITH CYCLES AND LATENT VARIABLES

BY STEPHAN BONGERS[1,*], PATRICK FORRÉ[1,†], JONAS PETERS[2] AND JORIS M. MOOIJ[3]

[1]*Informatics Institute, University of Amsterdam, [*]s.r.bongers@uva.nl; [†]p.d.forre@uva.nl*
[2]*Department of Mathematical Sciences, University of Copenhagen, jonas.peters@math.ku.dk*
[3]*Korteweg–De Vries Institute, University of Amsterdam, j.m.mooij@uva.nl*

Structural causal models (SCMs), also known as (nonparametric) structural equation models (SEMs), are widely used for causal modeling purposes. In particular, acyclic SCMs, also known as recursive SEMs, form a well-studied subclass of SCMs that generalize causal Bayesian networks to allow for latent confounders. In this paper, we investigate SCMs in a more general setting, allowing for the presence of both latent confounders and cycles. We show that in the presence of cycles, many of the convenient properties of acyclic SCMs do not hold in general: they do not always have a solution; they do not always induce unique observational, interventional and counterfactual distributions; a marginalization does not always exist, and if it exists the marginal model does not always respect the latent projection; they do not always satisfy a Markov property; and their graphs are not always consistent with their causal semantics. We prove that for SCMs in general each of these properties does hold under certain solvability conditions. Our work generalizes results for SCMs with cycles that were only known for certain special cases so far. We introduce the class of simple SCMs that extends the class of acyclic SCMs to the cyclic setting, while preserving many of the convenient properties of acyclic SCMs. With this paper, we aim to provide the foundations for a general theory of statistical causal modeling with SCMs.

1. Introduction. Structural causal models (SCMs), also known as (nonparametric) structural equation models (SEMs), are widely used for causal modeling purposes [4, 48, 51, 68]. They form the basis for many statistical methods that aim at inferring knowledge of the underlying causal structure from data (see, e.g., [7, 34, 42, 45, 52]). In these models, the causal relationships between the variables are expressed in the form of deterministic, functional relationships and probabilities are introduced through the assumption that certain variables are exogenous latent random variables. SCMs arose out of certain causal models that were first introduced in genetics [74], econometrics [22], electrical engineering [36, 37] and the social sciences [11, 21].

Acyclic SCMs, also known as recursive SEMs, form a special well-studied subclass of SCMs that generalize causal Bayesian networks [48]. They have many convenient properties (see, e.g., [14, 15, 31, 32, 47, 54, 73]): (i) they induce a unique distribution over the variables; (ii) they are closed under perfect interventions; (iii) they are closed under marginalizations; (iv) their marginalization respects the latent projection; (v) they obey (various equivalent versions of) the Markov property and (vi) their graphs express the causal relationships encoded by the SCM in an intuitive manner.

One important limitation of acyclic SCMs is that they cannot model systems that involve causal cycles. In many systems occurring in the real world, there are feedback loops between

Received May 2020; revised February 2021.
MSC2020 subject classifications. Primary 62A09, 68T30; secondary 68T37.
Key words and phrases. Structural causal models, causal graph, cycles, interventions, counterfactuals, solvability, Markov properties, marginalization.

observed variables. For example, in economics the price of a product may be a function of the demanded or supplied quantities, and vice versa, the demanded and supplied quantities may be functions of the price. The underlying dynamic processes describing such systems have an acyclic causal structure over time. However, causal cycles may arise when one approximates such systems over time [16, 39, 40] or when one describes the equilibrium states of these systems [3, 5, 24, 26, 30, 43, 53]. In particular, in [5] it was shown that the equilibrium states of a system governed by (random) differential equations can be described by an SCM that represents their causal semantics, which gives rise to a plethora of SCMs that include cycles (we provide some examples of such feedback systems in Appendix D.1 of the Supplementary Material [6]). In contrast to their acyclic counterparts, SCMs with cycles have enjoyed less attention in the literature and are not as well understood. In general, none of the above properties (i)–(vi) hold in the class of SCMs. However, some progress has been made in the case of discrete [46, 49] and linear models [24, 28, 59, 65–67], and more recently, for more general cyclic models the Markov properties have been elucidated [17].

Contributions. The purpose of this paper is to provide the foundations for a general theory of statistical causal modeling with SCMs. We study properties of SCMs and allow for cycles, latent variables and nonlinear functional relationships between the variables. We investigate to which extent and under which sufficient conditions each of the properties (i)–(vi) holds, in particular, in the presence of cycles. In the next paragraphs, we describe our contributions in more detail.

When there are cyclic functional relationships between variables, one encounters various technical complications, which even arise in the linear setting. The structural equations of an acyclic SCM trivially have a unique solution. This unique solvability property ensures that the SCM gives rise to a unique, well-defined probability distribution on the variables. In the case of cycles, however, this property may be violated, and consequently, the SCM may not have a solution at all, or may allow for multiple different probability distributions [23]. Even if one starts with a cyclic SCM that is uniquely solvable, performing an intervention on the SCM may lead to an intervened SCM that is not uniquely solvable. Hence, a cyclic SCM may not give rise to a unique, well-defined probability distribution corresponding to that intervention, and whether or not this happens may depend on the intervention. We provide sufficient conditions for the existence and uniqueness of these probability distributions after intervention. In general, it is not clear whether the solutions of the structural equations of an SCM are measurable if cycles are present. In addition, we provide sufficient and necessary conditions for the measurability of solution functions of cyclic SCMs.

SCMs provide a detailed modeling description of a system. Not all information may be necessary for a certain modeling task, which motivates to consider certain classes of SCMs to be equivalent. In this paper, we formally introduce several of such equivalence relations. For example, we consider two SCMs observationally equivalent if they cannot be distinguished based on observations alone. Observationally, equivalent SCMs can often still be distinguished by interventions. We consider two SCMs interventionally equivalent if they cannot be distinguished based on observations and interventions. While these concepts have been around in implicit form for acyclic SCMs, we formulate them in such a way that they also apply to cyclic SCMs that have either no solution at all or have multiple different induced probability distributions on the variables. Finally, we consider two SCMs counterfactually equivalent if they cannot be distinguished based on observations and interventions and in addition encode the same counterfactual distributions, which are the distributions induced by the so-called twin SCM via the twin network method [1]. These different equivalence relations formalize the different levels of abstraction in the so-called causal hierarchy [50, 64]. In addition, we add another, strong version of equivalence, such that equivalent SCMs have the same solutions. This notion clarifies ambiguities when a function is constant in one of its arguments, for example.

Marginalization becomes useful if not all variables are observed: given a joint probability distribution on some variables, we obtain a marginal distribution on a subset of the variables by integrating out the remaining variables. Analogously, we can marginalize an acyclic SCM by substituting the solutions of the structural equations of a subset of the endogenous variables into the structural equations of the remaining endogenous variables. For acyclic SCMs, the induced observational and interventional distributions of the marginalized SCM coincide with the marginals of the distributions induced by the original SCM (see [14, 15, 70, 73], a.o.). In other words, for acyclic SCMs the operation of marginalization preserves the probabilistic and causal semantics (restricted to the remaining variables). We show that for cyclic SCMs a marginalization does not always exist without further assumptions. In [17], it is shown that for modular SCMs, which can be seen as an SCM together with an additional structure of a compatible system of solution functions, a marginalization can be defined that preserves the probabilistic and causal semantics. We prove that this additional structure is not necessary and use a local unique solvability condition instead. Under this condition, we show that an SCM and its marginalization are observationally, interventionally and counterfactually equivalent on the remaining endogenous variables. Analogously, we define a marginalization operation on the associated graph of an SCM, which generalizes the latent projection [14, 71, 73]. In general, the marginalization of an SCM does not respect the latent projection of its associated graph, but we show that it does so under an additional local ancestral unique solvability condition.

In graphical models, Markov properties allow one to read off conditional independencies in a distribution directly from a graph. Various equivalent formulations of Markov properties exist for acyclic SCMs [31], one prominent example being the d-separation criterion, also known as the directed global Markov property, which was originally derived for Bayesian networks [47]. Markov properties have been of key importance to derive various central results regarding causal reasoning and causal discovery. For cyclic SCMs, however, the usual Markov properties do not hold in general, as was already pointed out by Spirtes [66]. His solution in terms of collapsed graphs was recently generalized and reformulated for a general class of causal graphical models [17] by adapting the notion of d-separation into what has been termed σ-separation. This resulted in a general directed global Markov property expressed in terms of σ-separation instead of d-separation. Here, we formulate these general Markov properties specifically within the framework of SCMs. Again, they only hold under certain unique solvability conditions.

In addition to its interpretation in terms of conditional independencies, the graph of an acyclic SCM also has a direct causal interpretation [48]. As was already observed in [46], the causal interpretation of SCMs with cycles can be counterintuitive, as the causal semantics under interventions no longer needs to be compatible with the structure imposed by the functional relations between the variables. We resolve this issue by showing that under certain ancestral unique solvability conditions the causal interpretation of SCMs is consistent with its graph.

Cycles lead to several technical complications related to solvability issues. We introduce a special subclass of (possibly cyclic) SCMs, the class of simple SCMs, for which most of these technical complications are absent and which preserves much of the simplicity of the theory for acyclic SCMs. A simple SCM is an SCM that is uniquely solvable with respect to every subset of the variables. Because of this strong solvability assumption, simple SCMs have all the convenient properties (i)–(vi): they always have uniquely defined observational, interventional and counterfactual distributions; we can perform every perfect intervention and marginalization on them and the result is again a simple SCM; marginalization does respect the latent projection; they obey the general directed global Markov property, and for special cases (including the acyclic, linear and discrete case) they obey the (stronger) directed global Markov property; their graphs have a direct and intuitive causal interpretation.

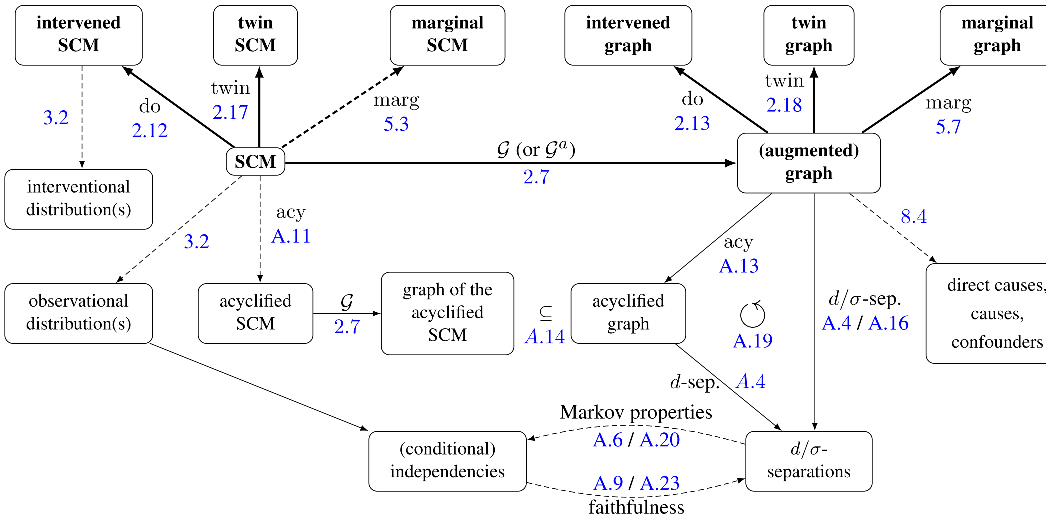

FIG. 1. *Overview of the objects constructed from an SCM and the mappings between them. The numbers correspond to the definition, proposition or theorem of the corresponding object, mapping or result. When an arrow is dashed, the relation only holds under nontrivial assumptions that can be found in the corresponding definition or theorem. The symbol "$\subseteq$" stands for the subgraph of a directed mixed graph (see Definition* A.1 *in the Supplementary Material* [6]) *and the symbol "↻" denotes that the surrounding diagram commutes. Table* 1 *gives an overview of the commutativity results for each pair of mappings between the objects with the names in bold.*

The scope of this paper is limited to establishing the foundations for statistical causal modeling with cyclic SCMs (Figure 3 in Appendix A.4 of the Supplementary Material [6] shows an overview of how SCMs relate to other causal graphical models). For a detailed discussion of causal reasoning, causal discovery and causal prediction with cyclic SCMs we refer the reader to other literature (e.g., [13, 20, 24, 25, 55, 57, 58]). Several recent results (generalizations of the do-calculus, adjustment criteria and an identification algorithm) for modular SCMs [18, 19] directly apply to the subclass of simple SCMs, as well. Finally, many causal discovery algorithms that have been designed for the acyclic case also apply to simple SCMs with no or only minor changes [41, 44].

Overview. Figure 1 gives an overview of the different objects that can be constructed from an SCM and the different mappings between them. For pairs of mappings between the objects with the names in bold, we prove commutativity results which are summarized in Table 1.

TABLE 1
Overview of the commutativity results of different pairs of mappings, defined on SCMs (left table) and on graphs (right table). All results apply under the assumptions stated in the corresponding proposition. The entries denoted by dots are omitted due to symmetry. We do not consider the commutativity of the twin operation with itself in this paper. Proposition 5.11 (*in parentheses*) *is not a commutativity result but a weaker relation. The graphical twin operator is only defined for directed graphs*

SCMs	do	twin	marg
$\mathcal{G}, \mathcal{G}^a$	2.14	2.19	(5.11)
do	2.15.(1)	2.21.(1)	5.5.(1)
twin	⋯	–	5.5.(2)
marg	⋯	⋯	5.4

Graphs	do	twin	marg
do	2.15.(1)	2.21.(2)	5.9.(1)
twin	⋯	–	5.9.(2)
marg	⋯	⋯	5.8

Outline. This paper is structured as follows: In Section 2, we provide a formal definition of SCMs and a natural notion of equivalence between SCMs, define the (augmented) graph corresponding to an SCM, and describe perfect interventions and counterfactuals. In Section 3, we discuss the concept of (unique) solvability, its properties and how it relates to self-cycles. In Section 4, we define and relate various equivalence relations between SCMs. In Section 5, we define a marginalization operation that is applicable to cyclic SCMs under certain conditions. We discuss several properties of this marginalization operation and discuss the relation with a marginalization operation defined on directed mixed graphs. In Section 6, we discuss Markov properties of SCMs. In Section 7, we discuss the causal interpretation of the graphs of SCMs. Section 8 introduces and discusses the class of simple SCMs.

The Supplementary Material [6] introduces causal graphical models in Appendix A. This section also contains details on Markov properties and modular SCMs. Appendix B provides additional (unique) solvability properties, some results for linear SCMs are discussed in Appendix C, other examples in Appendix D and the proofs of all the theoretical results are in Appendix E. Appendix F contains some lemmas and measurable selection theorems that are used in several proofs.

2. Structural causal models. In this section, we provide the definition and properties of structural causal models (SCMs). Our definition of SCMs slightly deviates from existing definitions [4, 48, 68], because we make the definition of the SCM independent of the random variables that solve it. This enables us to deal with the various technical complications that arise in the presence of cycles.

2.1. *Structural causal models and their solutions*.

DEFINITION 2.1 (Structural causal model). A *structural causal model (SCM)* is a tuple[1]

$$\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle,$$

where:

1. $\mathcal{I}$ is a finite index set of *endogenous variables*,
2. $\mathcal{J}$ is a disjoint finite index set of *exogenous variables*,
3. $\boldsymbol{\mathcal{X}} = \prod_{i\in\mathcal{I}} \mathcal{X}_i$ is the product of the domains of the endogenous variables, where each domain $\mathcal{X}_i$ is a standard measurable space (see Definition F.1),
4. $\boldsymbol{\mathcal{E}} = \prod_{j\in\mathcal{J}} \mathcal{E}_j$ is the product of the domains of the exogenous variables, where each domain $\mathcal{E}_j$ is a standard measurable space,
5. $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ is a measurable function that specifies the *causal mechanism*,
6. $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j\in\mathcal{J}} \mathbb{P}_{\mathcal{E}_j}$ is a product measure, the *exogenous distribution*, where $\mathbb{P}_{\mathcal{E}_j}$ is a probability measure on $\mathcal{E}_j$ for each $j \in \mathcal{J}$.[2]

In SCMs, the functional relationships between variables are expressed in terms of deterministic equations, where each equation expresses an endogenous variable (on the left-hand side) in terms of a causal mechanism depending on endogenous and exogenous variables (on the right-hand side). This allows us to model interventions in an unambiguous way by changing the causal mechanisms that target specific endogenous variables (see Section 2.4).

[1]We often use boldface for variables that have multiple components, for example, vectors in a Cartesian product.
[2]For the case $\mathcal{J} = \emptyset$, we have that $\boldsymbol{\mathcal{E}}$ is the singleton $\mathbf{1}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ is the degenerate probability measure $\mathbb{P}_{\mathbf{1}}$.

DEFINITION 2.2 (Structural equations). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call the set of equations

$$x_i = f_i(\boldsymbol{x}, \boldsymbol{e})\boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \quad \boldsymbol{e} \in \boldsymbol{\mathcal{E}}$$

for $i \in \mathcal{I}$ the *structural equations* of the structural causal model $\mathcal{M}$.

Although it is common to assume the absence of cyclic functional relations (see Definition 2.9), we make no such assumption here. In particular, we allow for self-cycles, which we will discuss in more detail in Sections 2.2 and 3.3.

The solutions of an SCM in terms of random variables are defined up to almost sure equality. Random variables that are almost surely equal are generally considered to be equivalent to each other for all practical purposes.

DEFINITION 2.3 (Solution). A pair $(\boldsymbol{X}, \boldsymbol{E})$ of random variables $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}$, $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$, where Ω is a probability space, is a *solution* of the SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ if:

1. $\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$, that is, the distribution of $\boldsymbol{E}$ is equal to $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$,[3] and
2. the *structural equations* are satisfied, that is,

$$\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E}) \quad \text{a.s.}$$

For convenience, we call a random variable $\boldsymbol{X}$ *a solution of* $\mathcal{M}$ if there exists a random variable $\boldsymbol{E}$ such that $(\boldsymbol{X}, \boldsymbol{E})$ forms a solution of $\mathcal{M}$.

Often, the endogenous random variables $\boldsymbol{X}$ can be observed, while the exogenous random variables $\boldsymbol{E}$ are treated as latent. Latent exogenous variables are often referred to as "disturbance terms" or "noise variables." For a solution $\boldsymbol{X}$, we call the distribution $\mathbb{P}^{\boldsymbol{X}}$ the *observational distribution of* $\mathcal{M}$ *associated to* $\boldsymbol{X}$. In general, there may be multiple different observational distributions associated to an SCM due to the existence of different solutions of the structural equations. This is a consequence of the allowance of cycles in SCMs, as the following simple example illustrates.

EXAMPLE 2.4 (Cyclic SCMs). For brevity, we use throughout this paper the notation $\boldsymbol{n} := \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$. Let $\mathcal{M} = \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ be an SCM[4] with $f_1(\boldsymbol{x}, e) = x_2$ and $f_2(\boldsymbol{x}, e) = x_1$, and $\mathbb{P}_{\mathbb{R}}$ an arbitrary probability measure on $\mathbb{R}$. Then (X, X) is a solution of $\mathcal{M}$ for any arbitrary random variable X with values in $\mathbb{R}$. Hence, any probability distribution on $\{(x, x) : x \in \mathbb{R}\}$ is an observational distribution associated to $\mathcal{M}$. Now consider instead the same SCM but with $f_1(\boldsymbol{x}, e) = x_2 + 1$. This SCM has no solutions at all, and hence induces no observational distribution.

Due to the fact that the structural equations only need to be satisfied almost surely, there may exist many different SCMs representing the same set of solutions (see Example D.4). It therefore seems natural not to differentiate between structural equations that have different solutions on at most a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set of exogenous variables. This leads to an equivalence relation between SCMs. To be able to state the equivalence relation concisely, we introduce the following notation: For subsets $\mathcal{U} \subseteq \mathcal{I}$ and $\mathcal{V} \subseteq \mathcal{J}$, we write $\boldsymbol{\mathcal{X}}_{\mathcal{U}} := \prod_{i \in \mathcal{U}} \mathcal{X}_i$ and $\boldsymbol{\mathcal{E}}_{\mathcal{V}} := \prod_{j \in \mathcal{V}} \mathcal{E}_j$. In particular, $\boldsymbol{\mathcal{X}}_{\varnothing}$ and $\boldsymbol{\mathcal{E}}_{\varnothing}$ are defined by the singleton $\mathbf{1}$. Moreover, for

[3] This implies that the components E_j of $\boldsymbol{E}$ are mutually independent, since $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$.

[4] We will abuse notation by using nondisjoint subsets of the natural numbers to index both endogenous and exogenous variables; these should be understood to be disjoint copies of the natural numbers: if we write $\mathcal{I} = \boldsymbol{n}$ and $\mathcal{J} = \boldsymbol{m}$, we mean instead $\mathcal{I} = \{1, 2, \dots, n\}$ and $\mathcal{J} = \{1', 2', \dots, m'\}$ where k' is a copy of k.

a subset $\mathcal{W} \subseteq \mathcal{I} \cup \mathcal{J}$, we use the convention that we write $\boldsymbol{\mathcal{X}}_{\mathcal{W}}$ and $\boldsymbol{\mathcal{E}}_{\mathcal{W}}$ instead of $\boldsymbol{\mathcal{X}}_{\mathcal{W}\cap\mathcal{I}}$ and $\boldsymbol{\mathcal{E}}_{\mathcal{W}\cap\mathcal{J}}$, respectively, and we adopt a similar notation for the (random) variables in those spaces, that is, we write $\boldsymbol{x}_{\mathcal{W}}$ and $\boldsymbol{e}_{\mathcal{W}}$ instead of $\boldsymbol{x}_{\mathcal{W}\cap\mathcal{I}}$ and $\boldsymbol{e}_{\mathcal{W}\cap\mathcal{J}}$, respectively. This allows us to define the following natural equivalence relation for SCMs.[5,6]

DEFINITION 2.5 (Equivalence). The two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ and $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ are *equivalent*, denoted by $\mathcal{M} \equiv \tilde{\mathcal{M}}$, if for all $i \in I$, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

$$x_i = f_i(\boldsymbol{x}, \boldsymbol{e}) \quad \iff \quad x_i = \tilde{f}_i(\boldsymbol{x}, \boldsymbol{e}).$$

Thus, two equivalent SCMs can only differ in terms of their causal mechanism. Importantly, equivalent SCMs have the same solutions and, as we will see in Sections 2.4 and 2.5, they have the same causal and counterfactual semantics (see Definitions 2.12 and 2.17, respectively). This equivalence relation on the set of all SCMs gives rise to the quotient set of equivalence classes of SCMs.

2.2. *The* (*augmented*) *graph*. We will now define two types of graphs that can be used for representing structural properties of the SCM. These graphical representations are related to Wright's path diagrams [74]. The structural properties of the functional relations between variables modeled by an SCM are specified by the causal mechanism of the SCM and can be encoded in an (augmented) graph. For the graphical notation and standard terminology on directed (mixed) graphs that is used throughout this paper, we refer the reader to Appendix A.1.

We first define the parents of an endogenous variable.

DEFINITION 2.6 (Parent). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ be an SCM. We call $k \in \mathcal{I} \cup \mathcal{J}$ a *parent of* $i \in \mathcal{I}$ if and only if there does not exist a measurable function[7] $\tilde{f}_i : \boldsymbol{\mathcal{X}}_{\setminus k} \times \boldsymbol{\mathcal{E}}_{\setminus k} \to \mathcal{X}_i$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

$$x_i = f_i(\boldsymbol{x}, \boldsymbol{e}) \quad \iff \quad x_i = \tilde{f}_i(\boldsymbol{x}_{\setminus k}, \boldsymbol{e}_{\setminus k}).$$

Exogenous variables have no parents by definition. These parental relations are preserved under the equivalence relation $\equiv$ on SCMs. They can be represented by a directed graph or a directed mixed graph.[8]

---

[5] An attempt at coarsening this notion of equivalence by replacing the quantifier "for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$" by "for almost every $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ under the observational distribution $\mathbb{P}^{\boldsymbol{X}}$" will not lead to a well-defined equivalence relation, since in general the observational distribution $\mathbb{P}^{\boldsymbol{X}}$ may be nonunique or even nonexistent. Refining it by replacing the quantifier "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" by "for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" would make it too fine for our purposes, since we assume the exogenous distribution to be fixed and we assume as usual that random variables that are almost surely identical are indistinguishable in practice. Note that the "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" and "for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$" quantifiers do not commute in general (see Example D.5)

[6] We may extend this definition to allow $\tilde{\mathcal{J}} \neq \mathcal{J}$ and for a larger class of SCMs such that the exogenous distribution does not factorize. Then, for any $\mathcal{M}$ that satisfies Definition 2.1, except for that it may have a nonfactorizing exogenous distribution, there exists an equivalent SCM with a factorizing exogenous distribution (and a different $\mathcal{J}$); the latter can be obtained by partitioning the exogenous components into independent tuples. This motivates why we can restrict ourselves in Definition 2.1 to factorizing exogenous distributions only. For some more discussion on the representation of latent confounders, see also Example D.6.

[7] For $\boldsymbol{\mathcal{X}} = \prod_{i\in\mathcal{I}} \mathcal{X}_i$, $\mathcal{I}$ some index set, $I \subseteq \mathcal{I}$ and $k \in \mathcal{I}$, we denote $\boldsymbol{\mathcal{X}}_{\setminus I} = \prod_{i\in\mathcal{I}\setminus I} \mathcal{X}_i$ and $\boldsymbol{\mathcal{X}}_{\setminus k} = \prod_{i\in\mathcal{I}\setminus\{k\}} \mathcal{X}_i$, and similarly for their elements.

[8] A *directed mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ consists of a set of nodes $\mathcal{V}$, a set of directed edges $\mathcal{E}$ and a set of bidirected edges $\mathcal{B}$ (see Definition A.1 for a more precise definition).

DEFINITION 2.7 (Graph and augmented graph). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We define:

1. the *augmented graph* $\mathcal{G}^a(\mathcal{M})$ as the directed graph with nodes $\mathcal{I} \cup \mathcal{J}$ and directed edges $u \to v$ if and only if $u \in \mathcal{I} \cup \mathcal{J}$ is a parent of $v \in \mathcal{I}$;
2. the *graph* $\mathcal{G}(\mathcal{M})$ as the directed mixed graph with nodes $\mathcal{I}$, directed edges $u \to v$ if and only if $u \in \mathcal{I}$ is a parent of $v \in \mathcal{I}$ and bidirected edges $u \leftrightarrow v$ if and only if there exists a $j \in \mathcal{J}$ that is a parent of both $u \in \mathcal{I}$ and $v \in \mathcal{I}$.

We call the mappings $\mathcal{G}^a$ and $\mathcal{G}$, that map $\mathcal{M}$ to $\mathcal{G}^a(\mathcal{M})$ and $\mathcal{G}(\mathcal{M})$, the *augmented graph mapping* and the *graph mapping*, respectively.

In particular, the augmented graph contains no directed edges pointing toward an exogenous variable, that is, $u \in \mathcal{I} \cup \mathcal{J}$ cannot be a parent of $v \in \mathcal{J}$, because they are not functionally related through the causal mechanism. We call a directed edge $i \to i$ in $\mathcal{G}^a(\mathcal{M})$ and $\mathcal{G}(\mathcal{M})$ (here, i is a parent of itself) a *self-cycle* at i. By definition, the mappings $\mathcal{G}^a$ and $\mathcal{G}$ are invariant under the equivalence relation $\equiv$ on SCMs, and hence the equivalence class of an SCM $\mathcal{M}$ is mapped to a unique augmented graph $\mathcal{G}^a(\mathcal{M})$ and a unique graph $\mathcal{G}(\mathcal{M})$.

EXAMPLE 2.8 (Graphs of an SCM). Let $\mathcal{M} = \langle \mathbf{5}, \mathbf{3}, \mathbb{R}^5, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$ be an SCM with causal mechanism given by

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = x_1 - x_1^2 + \alpha e_1^2, \qquad f_3(\boldsymbol{x}, \boldsymbol{e}) = -x_4 + e_2, \qquad f_5(\boldsymbol{x}, \boldsymbol{e}) = x_4 \cdot e_3,$$
$$f_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 + x_3 + x_4 + e_1, \qquad f_4(\boldsymbol{x}, \boldsymbol{e}) = x_2 + e_2,$$

where $\alpha \neq 0$ and $\mathbb{P}_{\mathbb{R}^3}$ is a product of three probability measures $\mathbb{P}_{\mathbb{R}}$ over $\mathbb{R}$ that are nondegenerate. The augmented graph $\mathcal{G}^a(\mathcal{M})$ and the graph $\mathcal{G}(\mathcal{M})$ of $\mathcal{M}$ are depicted[9] in Figure 2 (left and center). Observe that if α had been equal to zero, then the endogenous variable 1 would not have any parents in $\mathcal{G}^a(\mathcal{M})$, that is, it would not have a self-cycle and directed edge from any exogenous variables in $\mathcal{G}^a(\mathcal{M})$, and it would not have a self-cycle and bidirected edge from any other variable in $\mathcal{G}(\mathcal{M})$. Moreover, if one of the probability measures $\mathbb{P}_{\mathbb{R}}$ over $\mathbb{R}$ were degenerate, then some of the directed edges from the exogenous variables to the endogenous variables in the augmented graph $\mathcal{G}^a(\mathcal{M})$ and bidirected edges in the graph $\mathcal{G}(\mathcal{M})$ would be missing.

As is illustrated in this example, the augmented graph provides a more detailed representation than the graph. Therefore, we use the augmented graph as the standard graphical representation for SCMs, unless stated otherwise. For an SCM $\mathcal{M}$, we denote the sets $\mathrm{pa}_{\mathcal{G}^a(\mathcal{M})}(\mathcal{U})$, $\mathrm{ch}_{\mathcal{G}^a(\mathcal{M})}(\mathcal{U})$, $\mathrm{an}_{\mathcal{G}^a(\mathcal{M})}(\mathcal{U})$, etc., for some subset $\mathcal{U} \subseteq \mathcal{I} \cup \mathcal{J}$, by respectively $\mathrm{pa}(\mathcal{U})$, $\mathrm{ch}(\mathcal{U})$, $\mathrm{an}(\mathcal{U})$, etc., when the notation is clear from the context.

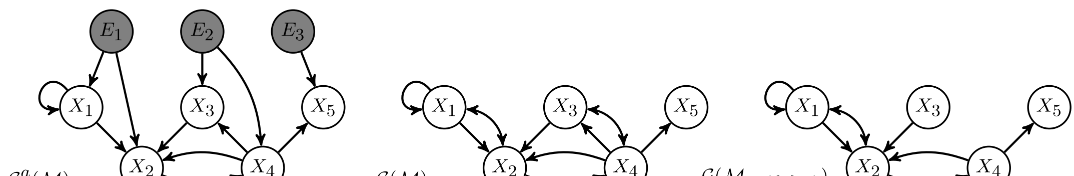

FIG. 2. *The augmented graph* (*left*) *and the graph* (*center*) *of the SCM $\mathcal{M}$ of Example* 2.8 *and the graph of the intervened SCM $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ of Example* 2.16 (*right*).

[9]For visualizing an (augmented) graph, we adapt the common convention of using random variables, with the index set as a subscript, instead of using the index set itself. With a slight abuse of notation, we still use the random variables notation in the (augmented) graph in the case that the SCM has no solution at all.

DEFINITION 2.9. We call an SCM $\mathcal{M}$ *acyclic* if $\mathcal{G}^a(\mathcal{M})$ is a directed acyclic graph (DAG). Otherwise, we call $\mathcal{M}$ *cyclic*.

Equivalently, an SCM $\mathcal{M}$ is acyclic if $\mathcal{G}(\mathcal{M})$ is an acyclic directed mixed graph (ADMG) [54]. Acyclic SCMs are also known as semi-Markovian SCMs [48, 71]. A commonly considered class of acyclic SCMs are the Markovian SCMs, which are acyclic SCMs for which each exogenous variable has at most one child. Several Markov properties were first shown for these models [32, 48, 71].

2.3. *Structurally minimal representations.* We have discussed an equivalence relation between SCMs in Section 2.1. In this subsection, we show that for each SCM there exists a representative of the equivalence class of that SCM for which each component of the causal mechanism does not depend on its nonparents (see also [51]).

DEFINITION 2.10 (Structurally minimal SCM). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$ *structurally minimal* if for all $i \in \mathcal{I}$ there exists a mapping $\tilde{f}_i : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(i)} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(i)} \to \mathcal{X}_i$ such that $f_i(\boldsymbol{x}, \boldsymbol{e}) = \tilde{f}_i(\boldsymbol{x}_{\mathrm{pa}(i)}, \boldsymbol{e}_{\mathrm{pa}(i)})$ for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.

We already encountered a structurally minimal SCM $\mathcal{M}$ in Example 2.8. Taking instead $\alpha = 0$ in that example gives an SCM $\mathcal{M}$ that is not structurally minimal, since the endogenous variable 1 is then not a parent of itself, while $f_1(\boldsymbol{x}, \boldsymbol{e})$ depends on x_1. However, the equivalent SCM where we have replaced the causal mechanism of 1 by $f_1(\boldsymbol{x}, \boldsymbol{e}) = 0$ yields a structurally minimal SCM. In general, there always exists an equivalent structurally minimal SCM.

PROPOSITION 2.11 (Existence of a structurally minimal SCM). *For an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, *there exists an equivalent SCM* $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *that is structurally minimal.*

For a causal mechanism $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ and a subset $\mathcal{U} \subseteq \mathcal{I}$, we write $\boldsymbol{f}_{\mathcal{U}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{U}}$ for the $\mathcal{U}$ components[10] of $\boldsymbol{f}$. A structurally minimal representation is compatible with the (augmented) graph, in the sense that for every $\mathcal{U} \subseteq \mathcal{I}$ there exists a unique measurable mapping $\tilde{\boldsymbol{f}}_{\mathcal{U}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{U})} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{U})} \to \boldsymbol{\mathcal{X}}_{\mathcal{U}}$ such that $\boldsymbol{f}_{\mathcal{U}}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_{\mathcal{U}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{U})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{U})})$ for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. Moreover, for any $\mathcal{U} \subseteq \mathcal{I}$ there exists a unique measurable mapping $\tilde{\boldsymbol{f}}_{\mathrm{an}(\mathcal{U})} : \boldsymbol{\mathcal{X}}_{\mathrm{an}(\mathcal{U})} \times \boldsymbol{\mathcal{E}}_{\mathrm{an}(\mathcal{U})} \to \boldsymbol{\mathcal{X}}_{\mathrm{an}(\mathcal{U})}$ with $\boldsymbol{f}_{\mathrm{an}(\mathcal{U})}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_{\mathcal{U}}(\boldsymbol{x}_{\mathrm{an}(\mathcal{U})}, \boldsymbol{e}_{\mathrm{an}(\mathcal{U})})$ for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.

2.4. *Interventions.* To define the causal semantics of SCMs, we consider here an idealized class of interventions introduced by Pearl [48] that we refer to as perfect interventions. Other types of interventions, like mechanism changes [72], fat-hand interventions [12], activity interventions [42] and stochastic versions of all these are at least as relevant, but we do not consider them here.

DEFINITION 2.12 (Perfect intervention on an SCM). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value. The *perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$ maps $\mathcal{M}$ to the SCM $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, where the *intervened causal mechanism* $\tilde{\boldsymbol{f}}$ is given by

$$\tilde{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \begin{cases} \xi_i, & i \in I, \\ f_i(\boldsymbol{x}, \boldsymbol{e}), & i \in \mathcal{I} \setminus I. \end{cases}$$

[10] For $\mathcal{U} = \varnothing$, we always consider the trivial mapping $\boldsymbol{f}_{\varnothing} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\varnothing}$ where $\boldsymbol{\mathcal{X}}_{\varnothing}$ is the singleton $\mathbf{1}$.

This operation $\mathrm{do}(I, \boldsymbol{\xi}_I)$ preserves the equivalence relation (see Definition 2.5) on the set of all SCMs, and hence this mapping induces a well-defined mapping on the set of equivalence classes of SCMs. Previous work has considered interventions only on a specific subset of endogenous variables [2, 3, 62]. Instead, we assume that we can intervene on any subset of endogenous variables in the model.

We define an analogous operation $\mathrm{do}(I)$ on directed mixed graphs.

DEFINITION 2.13 (Perfect intervention on a directed mixed graph). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $I \subseteq \mathcal{V}$ a subset. The perfect intervention $\mathrm{do}(I)$ maps $\mathcal{G}$ to the directed mixed graph $\mathrm{do}(I)(\mathcal{G}) := (\mathcal{V}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$, where $\tilde{\mathcal{E}} = \mathcal{E} \setminus \{v \to i : v \in \mathcal{V}, i \in I\}$ and $\tilde{\mathcal{B}} = \mathcal{B} \setminus \{v \leftrightarrow i : v \in \mathcal{V}, i \in \mathcal{I}\}$.

This operation simply removes all incoming edges on the nodes in I. The two notions of intervention are compatible with the (augmented) graph mapping.

PROPOSITION 2.14. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value. Then $(\mathcal{G}^a \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M}) = (\mathrm{do}(I) \circ \mathcal{G}^a)(\mathcal{M})$ and $(\mathcal{G} \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M}) = (\mathrm{do}(I) \circ \mathcal{G})(\mathcal{M})$.*

The two notions of perfect intervention satisfy the following elementary properties.

PROPOSITION 2.15. *For an SCM and a directed mixed graph*, *we have the following properties*:

1. *perfect interventions on disjoint subsets of variables commute*;
2. *acyclicity is preserved under perfect intervention.*

The following example shows that an SCM with a solution may not have a solution anymore after performing a perfect intervention on the SCM, and vice versa that an SCM without a solution may yield an SCM with a solution after intervention.

EXAMPLE 2.16 (Intervened SCM and its graphs). Consider the SCM $\mathcal{M}$ of Example 2.8 which has a solution if and only if $\alpha \geq 0$. Applying the perfect intervention $\mathrm{do}(\{3\}, 1)$ to $\mathcal{M}$ gives the intervened model $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ with the intervened causal mechanism

$$\tilde{f}_1(\boldsymbol{x}, \boldsymbol{e}) = x_1 - x_1^2 + \alpha e_1^2, \qquad \tilde{f}_3(\boldsymbol{x}, \boldsymbol{e}) = 1, \qquad \tilde{f}_5(\boldsymbol{x}, \boldsymbol{e}) = x_4 \cdot e_3,$$
$$\tilde{f}_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 + x_3 + x_4 + e_1, \qquad \tilde{f}_4(\boldsymbol{x}, \boldsymbol{e}) = x_2 + e_2,$$

for which the graph $\mathcal{G}(\mathcal{M}_{\mathrm{do}(\{3\},1)})$ is depicted in Figure 2 (right). This is an example where a perfect intervention leads to an intervened SCM $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ that does not have a solution anymore. In addition, performing a perfect intervention $\mathrm{do}(\{4\}, 1)$ on $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ yields again an SCM with a solution for $\alpha \geq 0$.

Recall that for each solution $\boldsymbol{X}$ of an SCM $\mathcal{M}$ we call the distribution $\mathbb{P}^{\boldsymbol{X}}$ the observational distribution of $\mathcal{M}$ associated to $\boldsymbol{X}$. For cyclic SCMs, the observational distribution is in general not unique.[11] For example, the SCM $\mathcal{M}$ of Example 2.8 has two different observational distributions if $\alpha > 0$. Similarly, an intervened SCM may induce a distribution that is

[11] In order to assure the existence of a unique observational distribution it is common to consider only SCMs for which the structural equations have a unique solution (see, e.g., Definition 7.1.1 in [48]). Although these SCMs induce a unique observational distribution, they generally do not induce a unique distribution after a perfect intervention.

not unique. Whenever the intervened SCM $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ has a solution $\boldsymbol{X}$ we therefore call the distribution $\mathbb{P}^{\boldsymbol{X}}$ the *interventional distribution of* $\mathcal{M}$ *under the perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *associated to* $\boldsymbol{X}$.[12]

2.5. *Counterfactuals*. The causal semantics of an SCM are described by the interventions on the SCM. Adding another layer of complexity, one can describe the counterfactual semantics of an SCM by the interventions on the so-called twin SCM, an idea introduced in [1].

DEFINITION 2.17 (Twin SCM). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. The *twin operation* maps $\mathcal{M}$ to the *twin structural causal model (twin SCM)*

$$\mathcal{M}^{\mathrm{twin}} := \langle \mathcal{I} \cup \mathcal{I}', \mathcal{J}, \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle,$$

where $\mathcal{I}' = \{i' : i \in \mathcal{I}\}$ is a copy of $\mathcal{I}$ and the causal mechanism $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}$ is the measurable function given by $\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{e}) = (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}), \boldsymbol{f}(\boldsymbol{x}', \boldsymbol{e}))$.

The twin operation on SCMs preserves the equivalence relation $\equiv$ on the set of all SCMs. We define an analogous twin operation $\mathrm{twin}(\mathcal{I})$ on directed graphs.

DEFINITION 2.18 (Twin graph). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph and $\mathcal{I} \subseteq \mathcal{V}$ a subset such that $\mathcal{J} := \mathcal{V} \setminus \mathcal{I}$ is *exogenous*, that is, $\mathrm{pa}_{\mathcal{G}}(\mathcal{J}) = \varnothing$. The $\mathrm{twin}(\mathcal{I})$ *operation* maps $\mathcal{G}$ to the *twin graph w.r.t.* $\mathcal{I}$ defined by $\mathrm{twin}(\mathcal{I})(\mathcal{G}) := (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, where:

1. $\tilde{\mathcal{V}} = \mathcal{V} \cup \mathcal{I}'$, where $\mathcal{I}'$ is a copy of $\mathcal{I}$,
2. $\tilde{\mathcal{E}} = \mathcal{E} \cup \mathcal{E}'$, where $\mathcal{E}'$ is given by

$$\mathcal{E}' = \{j \to i' : j \in \mathcal{J}, i \in \mathcal{I}, j \to i \in \mathcal{E}\} \cup \{\tilde{i}' \to i' : \tilde{i}, i \in \mathcal{I}, \tilde{i} \to i \in \mathcal{E}\}$$

with $i', \tilde{i}' \in \mathcal{I}'$ the respective copies of $i, \tilde{i} \in \mathcal{I}$.

Twin operations are compatible with the augmented graph mapping and preserve acyclicity.

PROPOSITION 2.19. *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be an SCM. Then* $(\mathcal{G}^a \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin}(\mathcal{I}) \circ \mathcal{G}^a)(\mathcal{M})$.

PROPOSITION 2.20. *For SCMs and directed graphs, we have that acyclicity is preserved under the twin operation.*

The perfect intervention and the twin operation for SCMs and directed graphs commute with each other in the following way.

PROPOSITION 2.21. *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be an SCM and* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *a directed graph. Then we have that perfect intervention commutes with the twin operation on both*:

1. *the SCM* $\mathcal{M}$: *for a subset* $I \subseteq \mathcal{I}$ *and value* $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, $(\mathrm{do}(I \cup I', \boldsymbol{\xi}_{I \cup I'})) \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin} \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M})$, *and*
2. *the directed graph* $\mathcal{G}$: *for subsets* $I \subseteq \mathcal{I} \subseteq \mathcal{V}$ *such that* $\mathcal{J} := \mathcal{V} \setminus \mathcal{I}$ *is exogenous*, $(\mathrm{do}(I \cup I') \circ \mathrm{twin}(\mathcal{I}))(\mathcal{G}) = (\mathrm{twin}(\mathcal{I}) \circ \mathrm{do}(I))(\mathcal{G})$,

where I' *is the copy of* I *in* $\mathcal{I}'$ *and* $\boldsymbol{\xi}_{I'} = \boldsymbol{\xi}_I$.

[12] In the literature, one often finds the notation $p(\boldsymbol{x})$ and $p(\boldsymbol{x} \mid \mathrm{do}(\boldsymbol{X}_I = \boldsymbol{x}_I))$ for the densities of the observational and interventional distribution, respectively, in case these are uniquely defined by the SCM (e.g., [48]).

Whenever the intervened twin SCM $(\mathcal{M}^{\text{twin}})_{\text{do}(\tilde{I},\boldsymbol{\xi}_{\tilde{I}})}$, where $\tilde{I} \subseteq \mathcal{I} \cup \mathcal{I}'$ and $\boldsymbol{\xi}_{\tilde{I}} \in \boldsymbol{\mathcal{X}}_{\tilde{I}}$, has a solution $(\boldsymbol{X}, \boldsymbol{X}')$, we call the distribution $\mathbb{P}^{(\boldsymbol{X},\boldsymbol{X}')}$ the *counterfactual distribution of* $\mathcal{M}$ *under the perfect intervention* $\text{do}(\tilde{I}, \boldsymbol{\xi}_{\tilde{I}})$ *associated to* $(\boldsymbol{X}, \boldsymbol{X}')$. In Example D.3, we provide an example of how counterfactuals can be sensibly formulated for a well-known market equilibrium model described in terms of a cyclic SCM.

The interpretation of counterfactual statements has received a lot of attention in the literature [1, 8, 33, 48, 61]. For acyclic graphs, an alternative graphical approach to counterfactuals is the framework of Single World Intervention Graphs (SWIGs) [60]. One topic of discussion is that there exist SCMs that induce the same observational and interventional distributions, but differ in their counterfactual statements [10] (see also Example D.7). This raises the question how one can estimate such SCMs from data.

3. Solvability. In this section, we introduce the notions of solvability and unique solvability with respect to a subset of the endogenous variables of an SCM. They describe the existence and uniqueness of measurable solution functions for the subsystem of structural equations that correspond with a certain subset of the endogenous variables. These notions play a central role in formulating sufficient conditions under which several properties of acyclic SCMs may be extended to the cyclic setting. For example, we show that solvability of an SCM is a sufficient and necessary condition for the existence of a solution of an SCM. Further, unique solvability of an SCM implies the uniqueness of the induced observational distribution.

3.1. *Definition of solvability.* Intuitively, one can think of the structural equations corresponding to a subset of endogenous variables $\mathcal{O} \subseteq \mathcal{I}$ as a description of how the subsystem formed by the variables $\mathcal{O}$ interacts with the rest of the system $\mathcal{I} \setminus \mathcal{O}$ through the variables $\text{pa}(\mathcal{O}) \setminus \mathcal{O}$. A solution function w.r.t. $\mathcal{O}$ assigns each input value $(\boldsymbol{x}_{\text{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\text{pa}(\mathcal{O})})$ of this subsystem to a specific output value $\boldsymbol{x}_{\mathcal{O}}$ of the subsystem. This is formalized as follows.

DEFINITION 3.1 (Solvability). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$ *solvable w.r.t.* $\mathcal{O} \subseteq \mathcal{I}$ if there exists a measurable mapping $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\text{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

$$\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{\text{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\text{pa}(\mathcal{O})}) \quad \Longrightarrow \quad \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e}).$$

We then call $\boldsymbol{g}_{\mathcal{O}}$ a *measurable solution function w.r.t.* $\mathcal{O}$ *for* $\mathcal{M}$. We call $\mathcal{M}$ *solvable* if it is solvable w.r.t. $\mathcal{I}$.

By definition, solvability w.r.t. a subset respects the equivalence relation $\equiv$ on SCMs. The measurable solution functions w.r.t. a certain subset do not always exist, and if they exist, they are not always uniquely defined. For example, for the SCM $\mathcal{M}$ in Example 2.8, the measurable solution functions w.r.t. $\{1\}$ are given by $g_1^{\pm}(e_1) = \pm\sqrt{\alpha e_1^2}$ if and only if $\alpha \geq 0$.

The following theorem states that various possible notions of "solvability" are equivalent.

THEOREM 3.2 (Sufficient and necessary conditions for solvability). *For an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, *the following are equivalent*:

1. $\mathcal{M}$ *has a solution* (*see Definition* 2.3);
2. *for* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-almost every* $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ *the structural equations* $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})$ *have a solution* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$;
3. $\mathcal{M}$ *is solvable* (*see Definition* 3.1).

While in the acyclic case, the above theorem is almost trivial, in the cyclic case the measure-theoretic aspects are not that obvious. In particular, to prove the existence of a *measurable* solution function $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ in case the structural equations have a solution for almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, we make use of a strong measurable selection theorem (see Theorem F.8 or [27]). This theorem implies that if there exists a solution $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}$, then there necessarily exists a random variable $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$ and a mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{g}(\boldsymbol{E}_{\text{pa}(\mathcal{I})})$ is a solution. However, it does not imply that there necessarily exists a random variable $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$ and a mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{E}_{\text{pa}(\mathcal{I})})$ holds a.s., for example, if $\boldsymbol{X}$ is a nontrivial mixture of such solutions (see Example D.8).

Solvability w.r.t. a strict subset of $\mathcal{I}$ is in general neither sufficient nor necessary for the existence of a (global) solution of the SCM. Consider, for example, the SCM $\mathcal{M}$ in Example 2.8 with $\alpha < 0$. Even though this SCM is solvable w.r.t. $\{2, 3, 4\}$, it is not (globally) solvable, and hence does not have any solution. In Proposition B.1, we provide a sufficient condition for solvability w.r.t. a strict subset of $\mathcal{I}$ that is similar to condition (2) in Theorem 3.2 in the sense that it is formulated in terms of the solutions of (a subset of) the structural equations without requiring measurability of the solutions. For the class of linear SCMs, we provide in Proposition C.2 a sufficient and necessary condition for solvability w.r.t. a subset of $\mathcal{I}$.

3.2. *Unique solvability.* The notion of unique solvability w.r.t. a subset $\mathcal{O} \subseteq \mathcal{I}$ is similar to the notion of solvability, but with the additional requirement that the measurable solution function $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\text{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ is unique up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set.

DEFINITION 3.3 (Unique solvability). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$ *uniquely solvable w.r.t.* $\mathcal{O} \subseteq \mathcal{I}$ if there exists a measurable mapping $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\text{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

$$\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{\text{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\text{pa}(\mathcal{O})}) \quad \Longleftrightarrow \quad \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e}).$$

We call $\mathcal{M}$ *uniquely solvable* if it is uniquely solvable w.r.t. $\mathcal{I}$.

If $\mathcal{M} \equiv \tilde{\mathcal{M}}$ and $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{O}$, then $\tilde{\mathcal{M}}$ is uniquely solvable w.r.t. $\mathcal{O}$, too, and the same mapping $\boldsymbol{g}_{\mathcal{O}}$ is a measurable solution function w.r.t. $\mathcal{O}$ for both $\mathcal{M}$ and $\tilde{\mathcal{M}}$.

The following result explains why the notions of (unique) solvability do not play an important role in the theory of acyclic SCMs.

PROPOSITION 3.4. *An acyclic SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is uniquely solvable w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$.*

We now illustrate that also cyclic SCMs can be uniquely solvable w.r.t. every subset.

EXAMPLE 3.5 (Cyclic SCM, uniquely solvable w.r.t. each subset). Consider the SCM $\mathcal{M} = \langle \mathbf{4}, \mathbf{4}, \mathbb{R}^4, \mathbb{R}^4, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^4} \rangle$ with causal mechanism given by

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = e_1, \qquad f_2(\boldsymbol{x}, \boldsymbol{e}) = e_2, \qquad f_3(\boldsymbol{x}, \boldsymbol{e}) = x_1 x_4 + e_3, \qquad f_4(\boldsymbol{x}, \boldsymbol{e}) = x_2 x_3 + e_4$$

and $\mathbb{P}_{\mathbb{R}^4}$ the standard-normal distribution on $\mathbb{R}^4$. This SCM $\mathcal{M}$ is uniquely solvable w.r.t. every subset and its (augmented) graph includes a cycle (see Figure 3).

FIG. 3. *Left*: *The graphs of the observationally equivalent SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ of Examples* 3.5 *and* 4.2, *respectively. Right*: *The graphs of the interventionally equivalent SCMs $\bar{\mathcal{M}}$ and $\hat{\mathcal{M}}$ of Example* 4.4.

Theorem 3.2 provides sufficient and necessary conditions for (global) solvability. The next theorem states that under the additional uniqueness requirement there exists a sufficient and necessary condition for unique solvability w.r.t. any subset (for solvability w.r.t. a subset we only have the sufficient condition provided in Proposition B.1), and moreover, that all solutions of a uniquely solvable SCM induce the same observational distribution.

THEOREM 3.6 (Sufficient and necessary conditions for unique solvability). *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM and $\mathcal{O} \subseteq \mathcal{I}$ a subset. The following are equivalent*:

1. *for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\setminus \mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\setminus \mathcal{O}}$ the structural equations*

$$\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})$$

have a unique solution $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$;

2. *$\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{O}$.*

Furthermore, if $\mathcal{M}$ is uniquely solvable, then there exists a solution, and all solutions have the same observational distribution.

It is well known that under acyclicity the observational distribution is unique. Theorem 3.6 generalizes this result to settings with cycles. For linear SCMs, the unique solvability condition w.r.t. a subset is equivalent to a matrix invertibility condition (see Proposition C.3).

In general, (unique) solvability w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply (unique) solvability w.r.t. a strict superset $\mathcal{O} \subsetneq \mathcal{V} \subseteq \mathcal{I}$ nor w.r.t. a strict subset $\mathcal{W} \subsetneq \mathcal{O}$ (see Example B.2). Moreover, (unique) solvability is in general not preserved under unions and intersections (see Appendix B.3).

3.3. *Self-cycles*. One can think of a structural equation of a single endogenous variable $i \in \mathcal{I}$ as describing a small subsystem that interacts with the rest of the system. If the output x_i of this subsystem is uniquely determined by the input $(\boldsymbol{x}_{\setminus i}, \boldsymbol{e})$ from the rest of the system (up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set), then i is not a parent of itself (see Definition 2.6).

PROPOSITION 3.7 (Self-cycles). *The SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is uniquely solvable w.r.t. $\{i\}$ for $i \in \mathcal{I}$ if and only if $\mathcal{G}^a(\mathcal{M})$ (or $\mathcal{G}(\mathcal{M})$) has no self-cycle $i \to i$ at $i \in \mathcal{I}$.*

A self-cycle at an endogenous variable denotes that that variable is not uniquely determined by its parents, up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. This implies that an SCM with a self-cycle at an endogenous variable in its graph can be either solvable, or not solvable, w.r.t. that variable. For the SCM $\mathcal{M}$ of Example 2.8, we have indeed that it is solvable w.r.t. $\{1\}$ for $\alpha > 0$, while for $\alpha < 0$ it is not. For linear SCMs with structural equations $X_i = \sum_{j \in \mathcal{I}} B_{ij} X_j + \sum_{k \in \mathcal{J}} \Gamma_{ik} E_k$, the endogenous variable $i \in \mathcal{I}$ has a self-cycle if and only if $B_{ii} = 1$ (see also Appendix C).

3.4. *Interventions*. The property of (unique) solvability is in general not preserved under perfect intervention. For example, a (uniquely) solvable SCM can lead to a nonuniquely solvable SCM after intervention, which either has no solution or has solutions with multiple induced distributions (see, e.g., Examples 2.16 and D.9). A sufficient condition for the intervened SCM to be (uniquely) solvable is that the original SCM has to be (uniquely) solvable w.r.t. the subset of nonintervened endogenous variables.

PROPOSITION 3.8. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is* (*uniquely*) *solvable w.r.t. $\mathcal{O} \subseteq \mathcal{I}$. Then, for any set I such that $\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O} \subseteq I \subseteq \mathcal{I} \setminus \mathcal{O}$ and value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ the intervened SCM $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}$ is* (*uniquely*) *solvable w.r.t. $\mathcal{O} \cup I$.*

Proposition 3.4 shows that acyclic SCMs are uniquely solvable w.r.t. every subset and hence are uniquely solvable after every perfect intervention. This also directly follows from the fact that acyclicity is preserved under perfect intervention (see Proposition 2.15). Moreover, since acyclicity is preserved under the twin operation (see Proposition 2.20), an acyclic SCM induces unique observational, interventional and counterfactual distributions.

3.5. *Ancestral* (*unique*) *solvability.* We saw that, in general, solvability w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply solvability w.r.t. a strict subset of $\mathcal{O}$. Here we show that it does imply solvability w.r.t. the ancestral subsets in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$, that is, in the induced subgraph of the graph $\mathcal{G}(\mathcal{M})$ on $\mathcal{O}$. A subset $\mathcal{A} \subseteq \mathcal{O}$ is called an *ancestral subset* in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$ if $\mathcal{A} = \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$, where $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$ are the ancestors of $\mathcal{A}$ according to the induced subgraph[13] $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$.

DEFINITION 3.9 (Ancestral (unique) solvability). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$ *ancestrally (uniquely) solvable w.r.t.* $\mathcal{O} \subseteq \mathcal{I}$ if $\mathcal{M}$ is (uniquely) solvable w.r.t. every ancestral subset in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$. We call $\mathcal{M}$ *ancestrally (uniquely) solvable* if it is ancestrally (uniquely) solvable w.r.t. $\mathcal{I}$.

PROPOSITION 3.10 (Solvability is equivalent to ancestral solvability). *The SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *is solvable w.r.t. the subset* $\mathcal{O} \subseteq \mathcal{I}$ *if and only if* $\mathcal{M}$ *is ancestrally solvable w.r.t.* $\mathcal{O}$.

A similar result does not hold for unique solvability. Although ancestral unique solvability w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ implies unique solvability w.r.t. $\mathcal{O}$, the converse does not hold in general, as the following example illustrates.

EXAMPLE 3.11 (Unique solvability w.r.t. $\mathcal{O}$ does not imply ancestral unique solvability w.r.t. $\mathcal{O}$). Consider the SCM $\mathcal{M} = \langle \mathbf{4}, \mathbf{1}, \mathbb{R}^4, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ with causal mechanism given by

$$f_1(\boldsymbol{x}, e) = e, \qquad f_2(\boldsymbol{x}, e) = x_2 \cdot \big(1 - \mathbf{1}_{\{0\}}(x_1 - x_3)\big) + 1,$$
$$f_3(\boldsymbol{x}, e) = x_3, \qquad f_4(\boldsymbol{x}, e) = x_3$$

and $\mathbb{P}_{\mathbb{R}}$ the standard-normal measure on $\mathbb{R}$. This SCM is uniquely solvable w.r.t. the set $\{2, 3\}$, and thus solvable w.r.t. this set. Although it is solvable w.r.t., the ancestral subset $\{3\}$ in $\mathcal{G}(\mathcal{M})_{\{2,3\}}$, depicted in Figure 4 (left), it is not uniquely solvable w.r.t. this subset, because the structural equation $x_3 = x_3$ holds for any $x_3 \in \mathbb{R}$. Hence, it is not ancestrally uniquely solvable w.r.t. $\{2, 3\}$.

However, for the class of linear SCMs we have that unique solvability w.r.t. $\mathcal{O}$ always implies ancestral unique solvability w.r.t. $\mathcal{O}$ (see Proposition C.4).

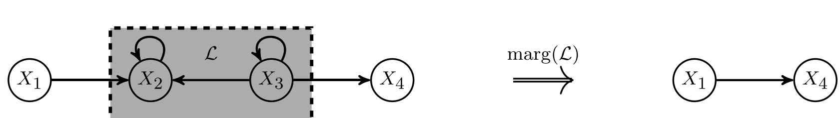

FIG. 4. *The graphs of the SCM* $\mathcal{M}$ (*left*) *of Example* 3.11 *and the marginal SCM* $\mathcal{M}_{\mathrm{marg}(\{2,3\})}$ (*right*) *of Example* 5.10.

[13]Here, one can also use the augmented graph $\mathcal{G}^a(\mathcal{M})$ on $\mathcal{O}$ since $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A}) = \mathrm{an}_{\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$ for every subset $\mathcal{A} \subseteq \mathcal{O} \subseteq \mathcal{I}$.

Although in general unique solvability is not preserved under unions, in Proposition B.4 we show that if an SCM is uniquely solvable w.r.t. two ancestral subsets and w.r.t. their intersection, then it is uniquely solvable w.r.t. their union. In general, the property of ancestral unique solvability is not preserved under perfect intervention, as can be seen in Example D.9. The notion of ancestral unique solvability will appear in various results in Sections 5 and 6.

4. Equivalences. In Section 2, we already encountered an equivalence relation on the class of SCMs (see Definition 2.5). The (augmented) graph of an SCM, its solutions and its induced observational, interventional and counterfactual distributions are preserved under this equivalence relation. In this section, we give several coarser equivalence relations on the class of SCMs: observational, interventional and counterfactual equivalence.

4.1. *Observational equivalence.* Observational equivalence is the property that two SCMs are indistinguishable on the basis of their observational distributions.

DEFINITION 4.1 (Observational equivalence). Two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ are *observationally equivalent w.r.t.* $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, denoted by $\mathcal{M} \equiv_{\text{obs}(\mathcal{O})} \tilde{\mathcal{M}}$, if $\boldsymbol{\mathcal{X}}_{\mathcal{O}} = \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{O}}$ and for all solutions $\boldsymbol{X}$ of $\mathcal{M}$ there exists a solution $\tilde{\boldsymbol{X}}$ of $\tilde{\mathcal{M}}$ such that $\mathbb{P}^{\boldsymbol{X}_{\mathcal{O}}} = \mathbb{P}^{\tilde{\boldsymbol{X}}_{\mathcal{O}}}$ and for all solutions $\tilde{\boldsymbol{X}}$ of $\tilde{\mathcal{M}}$ there exists a solution $\boldsymbol{X}$ of $\mathcal{M}$ such that $\mathbb{P}^{\boldsymbol{X}_{\mathcal{O}}} = \mathbb{P}^{\tilde{\boldsymbol{X}}_{\mathcal{O}}}$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are called *observationally equivalent* if they are observationally equivalent w.r.t. $\mathcal{I} = \tilde{\mathcal{I}}$.

Equivalent SCMs have the same solutions, and hence they are observationally equivalent w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$. However, observational equivalence does not imply equivalence.

EXAMPLE 4.2 (Observational equivalence does not imply equivalence). Consider the SCM $\tilde{\mathcal{M}}$ that is the same as $\mathcal{M}$ of Example 3.5 but with the causal mechanism $\tilde{\boldsymbol{f}}$ given by

$$\tilde{f}_1(\boldsymbol{x}, \boldsymbol{e}) := e_1, \qquad \tilde{f}_2(\boldsymbol{x}, \boldsymbol{e}) := e_2, \qquad \tilde{f}_3(\boldsymbol{x}, \boldsymbol{e}) := \frac{x_1 e_4 + e_3}{1 - x_1 x_2}, \qquad \tilde{f}_4(\boldsymbol{x}, \boldsymbol{e}) := \frac{x_2 e_3 + e_4}{1 - x_1 x_2}.$$

This SCM $\tilde{\mathcal{M}}$ is observationally equivalent to the SCM $\mathcal{M}$. Because both SCMs have a different (augmented) graph they are not equivalent to each other (see Figure 3).

This example shows that if two SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are observationally equivalent, then their associated augmented graphs $\mathcal{G}^a(\mathcal{M})$ and $\mathcal{G}^a(\tilde{\mathcal{M}})$ are not necessarily equal to each other.

4.2. *Interventional equivalence.* We consider two SCMs to be interventionally equivalent if they induce the same interventional distributions under all perfect interventions.

DEFINITION 4.3 (Interventional equivalence). Two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ are *interventionally equivalent w.r.t.* $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, denoted by $\mathcal{M} \equiv_{\text{int}(\mathcal{O})} \tilde{\mathcal{M}}$, if $\boldsymbol{\mathcal{X}}_{\mathcal{O}} = \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{O}}$ and for every $I \subseteq \mathcal{O}$ and every value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ their intervened models $\mathcal{M}_{\text{do}(I, \boldsymbol{\xi}_I)}$ and $\tilde{\mathcal{M}}_{\text{do}(I, \boldsymbol{\xi}_I)}$ are observationally equivalent with respect to $\mathcal{O}$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are called *interventionally equivalent* if they are interventionally equivalent w.r.t. $\mathcal{I} = \tilde{\mathcal{I}}$.

Equivalent SCMs have the same solutions under every perfect intervention, and hence they are interventionally equivalent w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$. SCMs that are interventionally equivalent w.r.t. a subset $\mathcal{O} \subseteq \mathcal{I}$ are interventionally equivalent w.r.t. every strict subset $\mathcal{W} \subsetneq \mathcal{O}$. But in general, they are not interventionally equivalent w.r.t. a strict superset $\mathcal{O} \subsetneq \mathcal{V} \subseteq \mathcal{I}$, as can be seen in Example 4.2, where the SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent w.r.t. $\{1, 2\}$ but are not interventionally equivalent. Interventional equivalence w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ implies observational equivalence w.r.t. $\mathcal{O}$, since the empty perfect intervention ($I = \varnothing$) is a special case of a perfect intervention. However, observational equivalence w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply interventional equivalence w.r.t. $\mathcal{O}$ in general, as can be seen in Example 4.2, where the SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are observationally equivalent but not interventionally equivalent.

Although interventional equivalence is a finer notion than observational equivalence, we have that if two SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent, then their associated augmented graphs $\mathcal{G}^a(\mathcal{M})$ and $\mathcal{G}^a(\tilde{\mathcal{M}})$ are not necessarily equal to each other.

EXAMPLE 4.4 (Interventionally equivalent SCMs with different graphs). Consider the SCM $\bar{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \{-1, 1\}^2, \{-1, 1\}^2, \bar{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and the SCM $\hat{\mathcal{M}}$ that is the same as $\bar{\mathcal{M}}$ except for its causal mechanism $\hat{\boldsymbol{f}}$, where the causal mechanisms are given by

$$\bar{f}_1(\boldsymbol{x}, \boldsymbol{e}) = e_1, \qquad \bar{f}_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 e_2, \qquad \hat{f}_1(\boldsymbol{x}, \boldsymbol{e}) = e_1, \qquad \hat{f}_2(\boldsymbol{x}, \boldsymbol{e}) = e_2,$$

and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^{\boldsymbol{E}}$ with $E_1, E_2 \sim \mathcal{U}(\{-1, 1\})$ uniformly distributed and $E_1 \perp\!\!\!\perp E_2$. Then $\bar{\mathcal{M}}$ and $\hat{\mathcal{M}}$ are interventionally equivalent although $\mathcal{G}(\bar{\mathcal{M}})$ is not equal to $\mathcal{G}(\hat{\mathcal{M}})$ (see Figure 3).

Example D.6 showcases an SCM with two endogenous and three exogenous variables, for which there is no interventionally equivalent SCM (satisfying smoothness constraints) with one exogenous variable taking values in $\mathbb{R}^2$ whose first and second components enter in the first and second structural equation, respectively. In this sense, representing confounders with dependent exogenous variables can be nontrivial in nonlinear models.

4.3. *Counterfactual equivalence.* We consider two SCMs to be counterfactually equivalent if their twin SCMs induce the same counterfactual distributions under every perfect intervention.

DEFINITION 4.5 (Counterfactual equivalence). Two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ are *counterfactually equivalent with respect to* $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, denoted by $\mathcal{M} \equiv_{\mathrm{cf}(\mathcal{O})} \tilde{\mathcal{M}}$, if $\mathcal{M}^{\mathrm{twin}}$ and $\tilde{\mathcal{M}}^{\mathrm{twin}}$ are interventionally equivalent with respect to $\mathcal{O} \cup \mathcal{O}'$, where $\mathcal{O}'$ corresponds to the copy of $\mathcal{O}$ in $\mathcal{I}' \cap \tilde{\mathcal{I}}'$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are called *counterfactually equivalent* if they are counterfactually equivalent with respect to $\mathcal{I} = \tilde{\mathcal{I}}$.

The notion of counterfactual equivalence is coarser than equivalence and finer than interventional equivalence.

PROPOSITION 4.6. *For SCMs, we have that equivalence implies counterfactual equivalence w.r.t.* $\mathcal{O}$, *which in turn implies interventional equivalence w.r.t.* $\mathcal{O}$, *for any* $\mathcal{O} \subseteq \mathcal{I}$.

Interventionally equivalent SCMs that have the same causal mechanism (that differ only in their exogenous distribution) may not be counterfactually equivalent (see, e.g., Example D.7). Although the notion of counterfactual equivalence is finer than the notion of observational and interventional equivalence, the (augmented) graphs for counterfactually equivalent SCMs are in general not equal to each other (see Example D.10).

4.4. *Relations between equivalences.* The definitions of observational, interventional and counterfactual equivalence provide equivalence relations on the set of all SCMs. For two SCMs to be observationally, interventionally or counterfactually equivalent w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, the domains of their endogenous variables $\mathcal{O}$ have to be equal, that is, $\boldsymbol{\mathcal{X}}_{\mathcal{O}} = \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{O}}$. Apart from that, the index sets of the endogenous and the exogenous variables, the spaces of the other endogenous and exogenous variables, the causal mechanism and the exogenous probability measure may all differ. The observational, interventional and counterfactual equivalence classes w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$ are related in the following way (see Proposition 4.6):

$$\begin{aligned}
\mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are equivalent} \quad &\Longrightarrow \quad \mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are counterfactually equivalent w.r.t. } \mathcal{O} \\
&\Longrightarrow \quad \mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are interventionally equivalent w.r.t. } \mathcal{O} \\
&\Longrightarrow \quad \mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are observationally equivalent w.r.t. } \mathcal{O}.
\end{aligned}$$

This hierarchy allows us to compare SCMs at different levels of abstraction and formally establishes the "ladder" of causation (last two implications) [48, 50, 64].

5. Marginalizations. In this section, we show how, and under which condition, one can marginalize an SCM over a subset $\mathcal{L} \subseteq \mathcal{I}$ of endogenous variables (thereby "hiding" the variables $\mathcal{L}$), to another SCM on the margin $\mathcal{I} \setminus \mathcal{L}$ that is observationally, interventionally and even counterfactually equivalent with respect to $\mathcal{I} \setminus \mathcal{L}$. In other words, we provide a formal notion of marginalization and show that this preserves the probabilistic, causal and counterfactual semantics on the margin.

The problem of marginalization of directed graphical models has been addressed for acyclic graph structures, for example, ADMGs and mDAGs (see [14, 15, 54, 56, 73], a.o.), and more recently in [17] for certain graph structures ("HEDGes") that may include cycles. Although in the acyclic setting it has been shown that the marginalization for some of these graph structures preserves the probabilistic and causal semantics, in the cyclic setting this has only been shown for modular SCMs [17]. We show that without the additional structure of a compatible system of solution functions (see Appendix A.3) one can still define a marginalization for SCMs under certain local unique solvability conditions. Intuitively, the idea is that if the state of a subsystem of endogenous variables is uniquely determined by the parents outside of this subsystem, then one can ignore the internals of this subsystem by treating it as a "black box" that can be described by certain measurable solution functions (see Figure 4). One can marginalize over this subsystem by substituting these measurable solution functions into the rest of the model, thereby removing the functional dependencies on the variables of the subsystem from the rest of the system, while preserving the probabilistic, causal and the counterfactual semantics of the rest of the system. We show that in general this marginalization operation defined on SCMs does not respect the latent projection on its associated (augmented) graph, where the latent projection is a similar marginalization operation defined on directed mixed graphs [14, 71, 73]. We show that under certain stronger local ancestral unique solvability conditions the marginalization does respect the latent projection.

5.1. *Marginalization of a structural causal model.* Before we show how one can marginalize an SCM w.r.t. a subset of endogenous variables, we first point out that in general it is not always possible to find an SCM on the margin that preserves the causal semantics, as the following example illustrates.

EXAMPLE 5.1 (No SCM on the margin preserves the causal semantics). Consider the SCM $\mathcal{M} = \langle \mathbf{3}, \varnothing, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ with causal mechanism $f_1(\boldsymbol{x}) = x_1 + x_2 + x_3$, $f_2(\boldsymbol{x}) = x_2$,

$f_3(\boldsymbol{x}) = 0$. Then there exists no SCM $\tilde{\mathcal{M}}$ on the endogenous variables $\{2, 3\}$ that is interventionally equivalent to $\mathcal{M}$ w.r.t. $\{2, 3\}$. To see this, suppose there exists such an SCM $\tilde{\mathcal{M}}$, then for every $(\xi_2, \xi_3) \in \boldsymbol{\mathcal{X}}_{\{2,3\}}$ such that $\xi_2 + \xi_3 \neq 0$ the intervened model $\tilde{\mathcal{M}}_{\mathrm{do}(\{2,3\},(\xi_2,\xi_3))}$ has a solution but $\mathcal{M}_{\mathrm{do}(\{2,3\},(\xi_2,\xi_3))}$ does not.

More generally, for an SCM $\mathcal{M}$ that is not solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$ there is no SCM $\tilde{\mathcal{M}}$ on the endogenous variables $\mathcal{I} \setminus \mathcal{L}$ that is interventionally equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.

The following example illustrates that for an SCM that is uniquely solvable w.r.t. a subset there exists an SCM on the margin that preserves the causal semantics.

EXAMPLE 5.2 (SCM on the margin that preserves the causal semantics). Consider the SCM $\mathcal{M}$ of Example 3.11 that is uniquely solvable w.r.t. the subset $\mathcal{L} = \{2, 3\}$ (depicted by the gray box in Figure 4). Substituting the measurable solution functions $\boldsymbol{g}_{\mathcal{L}}$ into the causal mechanism components f_1 and f_4 for the remaining endogenous variables $\{1, 4\}$ gives a "marginal" causal mechanism $\tilde{f}_1(\boldsymbol{x}, e) := e$ and $\tilde{f}_4(\boldsymbol{x}, e) := x_1$. This defines an SCM $\tilde{\mathcal{M}}$ on the margin $\mathcal{I} \setminus \mathcal{L} = \{1, 4\}$ that is interventionally equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$ to $\mathcal{M}$.

In general, for an SCM $\mathcal{M}$ and a given subset $\mathcal{L} \subseteq \mathcal{I}$ of endogenous variables and its complement $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$, we can consider the "subsystem" of structural equations $\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e})$. If $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ with measurable solution function $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}}$, then for each input $(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}) \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})}$ of the subsystem, there exists an output $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}$, which is unique for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})}}$-almost every $\boldsymbol{e}_{\mathrm{pa}(\mathcal{L})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})}$ and for all $\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}$. We can remove this subsystem of endogenous variables from the model by substitution. This leads to a marginal SCM that is observationally, interventionally and counterfactually equivalent to the original SCM w.r.t. the margin, as we prove in Theorem 5.6.

DEFINITION 5.3 (Marginalization of an SCM). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$ and let $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$. For $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})} \to \boldsymbol{\mathcal{L}}$, any measurable solution function of $\mathcal{M}$ w.r.t. $\mathcal{L}$, we call the SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L})} := \langle \mathcal{O}, \mathcal{J}, \boldsymbol{\mathcal{X}}_{\mathcal{O}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with the *marginal causal mechanism* $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ given by

$$\tilde{\boldsymbol{f}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = \boldsymbol{f}_{\mathcal{O}}\big(\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}), \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}\big),$$

a *marginalization of* $\mathcal{M}$ *w.r.t.* $\mathcal{L}$. We denote by $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ the equivalence class of the marginalizations of $\mathcal{M}$ w.r.t. $\mathcal{L}$.

The marginalization of $\mathcal{M}$ w.r.t. $\mathcal{L}$ is defined up to the equivalence $\equiv$ on SCMs, since the measurable solution functions $\boldsymbol{g}_{\mathcal{L}}$ are uniquely defined up to $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null sets. With this definition at hand, we can always construct a marginal SCM over a subset of the endogenous variables of an acyclic SCM by mere substitution (see also Proposition 3.4). Moreover, this definition extends that notion to SCMs that are uniquely solvable w.r.t. a certain subset. For linear SCMs this condition translates into a matrix invertibility condition, and since substitution preserves linearity, marginalization yields a linear marginal SCM (see Proposition C.5).

In general, marginalization is not always defined for all subsets. For instance, the SCM of Example 3.11 cannot be marginalized over the variable 3 (due to the self-cycle at 3), but can be marginalized over the variables 2 and 3 together. It follows from Proposition 3.7 that we can only marginalize over a single variable if that variable has no self-cycle. Note that we may introduce new self-cycles if we marginalize over a subset of variables, as can be seen, for example, from the SCM $\mathcal{M}$ in Example 2.8. This SCM has only one self-cycle; however, marginalizing w.r.t. $\{2\}$ gives a marginal SCM with another self-cycle at variable 4.

The definition of marginalization satisfies an intuitive property: if we can marginalize over two disjoint subsets after each other, then we can also marginalize over the union of those subsets at once, and the respective results agree.

PROPOSITION 5.4. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L}_1 \subseteq \mathcal{I}$ and let $\mathcal{L}_2 \subseteq \mathcal{I}$ be a subset disjoint from $\mathcal{L}_1$. Then $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$ if and only if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1 \cup \mathcal{L}_2$, Moreover,* $\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1)(\mathcal{M}) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{M})$.

In this proposition, $\mathcal{L}_1$ and $\mathcal{L}_2$ have to be disjoint, since marginalizing first over $\mathcal{L}_1$ gives a marginal SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ with endogenous variables $\mathcal{I} \setminus \mathcal{L}_1$.

Next, we show that the distributions of a marginal SCM are identical to the marginal distributions induced by the original SCM. A simple proof of this result proceeds by showing that both the intervention and the twin operation commute with marginalization.

PROPOSITION 5.5. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$. Then the marginalization* $\mathrm{marg}(\mathcal{L})$ *commutes with both*:

1. *the perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *for a subset $I \subseteq \mathcal{I} \setminus \mathcal{L}$ and a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, that is,* $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M}) = (\mathrm{do}(I, \boldsymbol{\xi}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$, *and*
2. *the twin operation* twin, *that is,* $(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin} \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$,

where $\mathcal{L}'$ is the copy of $\mathcal{L}$ in $\mathcal{I}'$.

With Proposition 5.5 at hand, we can prove the main result of this subsection.

THEOREM 5.6 (Marginalization of an SCM preserves the observational, causal and counterfactual semantics). *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$. Then $\mathcal{M}$ and* $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ *are observationally, interventionally and counterfactually equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.*

This shows that our definition of marginalization (Definition 5.3) preserves the probabilistic, causal and counterfactual semantics, under a certain local unique solvability condition. Moreover, this allows us to marginalize SCMs w.r.t. a certain subset that do not satisfy the additional assumptions imposed by modular SCMs, for example, the SCM $\mathcal{M}$ of Example 3.11 does not have any additional structure of a compatible system of solution functions, but $\mathcal{M}$ can be marginalized w.r.t. the subset $\{2, 3\}$ (see Appendix A.3).

In general, interventional equivalence does not imply counterfactual equivalence (see, e.g., Example D.7). However, for our definition of marginalization we arrive at a marginal SCM that is not only interventionally equivalent, but also counterfactually equivalent w.r.t. the margin.

For an SCM $\mathcal{M}$, unique solvability w.r.t. a certain subset $\mathcal{L} \subseteq \mathcal{I}$ is a sufficient, but not a necessary condition for the existence of an SCM $\tilde{\mathcal{M}}$ on the margin $\mathcal{I} \setminus \mathcal{L}$ such that $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are counterfactually equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$ (see, e.g., Example D.11). Hence, in certain cases it may be possible to relax the uniqueness condition.

5.2. *Marginalization of a graph.* We now turn to a marginalization operation for directed mixed graphs, which we call the latent projection. This name is inspired from a similar construction on directed mixed graphs in [73]. In [73], the authors concentrate on a mapping between directed mixed graphs and show that it preserves conditional independence properties (see also [71]). In this subsection, we provide a sufficient condition for the marginalization of an SCM to respect the latent projection, that is, that the augmented graph of the marginal SCM is a subgraph of the latent projection of the augmented graph of the original SCM.

DEFINITION 5.7 (Marginalization/latent projection of a directed mixed graph). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathcal{L} \subseteq \mathcal{V}$ a subset. The *marginalization of* $\mathcal{G}$ *w.r.t.* $\mathcal{L}$ or the *latent projection of* $\mathcal{G}$ *onto* $\mathcal{V} \setminus \mathcal{L}$ maps $\mathcal{G}$ to the *marginal graph* $\mathrm{marg}(\mathcal{L})(\mathcal{G}) := (\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$, where:

1. $\tilde{\mathcal{V}} = \mathcal{V} \setminus \mathcal{L}$,
2. for $i, j \in \tilde{\mathcal{V}}$: $i \to j \in \tilde{\mathcal{E}}$ if and only if there exists a directed path $i \to \ell_1 \to \cdots \to \ell_n \to j$ in $\mathcal{G}$ with $n \geq 0$ and $\ell_1, \dots, \ell_n \in \mathcal{L}$,
3. for $i \neq j \in \tilde{\mathcal{V}}$: $i \leftrightarrow j \in \tilde{\mathcal{B}}$ if and only if:

   (a) there exist $n, m \geq 0$, $\ell_1, \dots, \ell_n \in \mathcal{L}$, $\tilde{\ell}_1, \dots, \tilde{\ell}_m \in \mathcal{L}$ such that $i \leftarrow l_1 \leftarrow l_2 \leftarrow \cdots \leftarrow \ell_n \leftrightarrow \tilde{\ell}_m \to \tilde{\ell}_{m-1} \to \cdots \to \tilde{\ell}_1 \to j$ in $\mathcal{G}$, or

   (b) there exist $n, m \geq 1$, $\ell_1, \dots, \ell_n \in \mathcal{L}$, $\tilde{\ell}_1, \dots, \tilde{\ell}_m \in \mathcal{L}$ such that $i \leftarrow l_1 \leftarrow l_2 \leftarrow \cdots \leftarrow \ell_n$ and $\tilde{\ell}_m \to \tilde{\ell}_{m-1} \to \cdots \to \tilde{\ell}_1 \to j$ in $\mathcal{G}$ and $\ell_n = \tilde{\ell}_m$.

Note that this gives $\mathcal{G}(\mathcal{M}) = \mathrm{marg}(\mathcal{J})(\mathcal{G}^a(\mathcal{M}))$ for any SCM $\mathcal{M}$. Further, for a subgraph $\mathcal{H} \subseteq \mathcal{G}$ we have $\mathrm{marg}(\mathcal{L})(\mathcal{H}) \subseteq \mathrm{marg}(\mathcal{L})(\mathcal{G})$ for any subset of nodes $\mathcal{L}$. It does not matter in which order we project out the nodes or if we perform several projections at once.

PROPOSITION 5.8. *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *be a directed mixed graph and* $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{V}$ *two disjoint subsets. Then* $(\mathrm{marg}(\mathcal{L}_1) \circ \mathrm{marg}(\mathcal{L}_2))(\mathcal{G}) = (\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1))(\mathcal{G}) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{G})$.

Similar to the definition of marginalization for SCMs, this definition of the latent projection commutes with both the (graphical) perfect intervention and the twin operation.

PROPOSITION 5.9. *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *be a directed mixed graph and* $\mathcal{L}, \mathcal{I}, I \subseteq \mathcal{V}$ *subsets. Then the marginalization* $\mathrm{marg}(\mathcal{L})$ *commutes with both*:

1. *perfect intervention* $\mathrm{do}(I)$ *if* I *is disjoint from* $\mathcal{L}$, *that is*, $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I))(\mathcal{G}) = (\mathrm{do}(I) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})$, *and*
2. *the twin operation* $\mathrm{twin}(\mathcal{I})$ *if* $\mathcal{B} = \varnothing$, $\mathcal{J} := \mathcal{V} \setminus \mathcal{I}$ *is exogenous* (*i.e.*, $\mathrm{pa}_{\mathcal{G}}(\mathcal{J}) = \varnothing$) *and* $\mathcal{L} \subseteq \mathcal{I}$, *that is*, $(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin}(\mathcal{I}))(\mathcal{G}) = (\mathrm{twin}(\mathcal{I} \setminus \mathcal{L}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})$,

where $\mathcal{L}'$ *is the copy of* $\mathcal{L}$ *in* $\mathcal{I}'$.

An example of an SCM for which a marginalization respects the latent projection is the SCM $\mathcal{M}$ of Example 2.8. Marginalizing $\mathcal{M}$ w.r.t. $\mathcal{L} = \{2\}$ gives a marginal SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ with a graph that is a subgraph of the latent projection of the graph of the SCM $\mathcal{M}$ onto $\mathcal{I} \setminus \mathcal{L}$. In general, not all marginalizations respect the latent projection, as is illustrated in the following example.

EXAMPLE 5.10 (Marginalization does not respect the latent projection). Consider the SCM $\mathcal{M}$ of Example 3.11. Although $\mathcal{M}$ and its marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ with $\mathcal{L} = \{2, 3\}$ are interventionally equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L} = \{1, 4\}$, the graph $\mathcal{G}(\mathcal{M}_{\mathrm{marg}(\mathcal{L})})$ is not a subgraph of the latent projection of $\mathcal{G}(\mathcal{M})$ onto $\mathcal{I} \setminus \mathcal{L}$, as can be verified from the graphs depicted in Figure 4.

Under the local ancestral unique solvability condition, which is a stronger condition than the local unique solvability condition (i.e., ancestral unique solvability w.r.t. a subset implies unique solvability w.r.t. that subset), one can prove that the marginalization of an SCM respects the latent projection.

PROPOSITION 5.11. *Let $\mathcal{M}$ be an SCM that is ancestrally uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$. Then* $(\mathcal{G}^a \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M}) \subseteq (\mathrm{marg}(\mathcal{L}) \circ \mathcal{G}^a)(\mathcal{M})$ *and* $(\mathcal{G} \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M}) \subseteq (\mathrm{marg}(\mathcal{L}) \circ \mathcal{G})(\mathcal{M})$.

The (augmented) graph of a marginalized SCM can be a strict subgraph of the corresponding latent projection if, for example, certain paths cancel each other out after the substitution of the measurable solution function(s) into the causal mechanism(s) on the margin (see Example D.12). For acyclic SCMs, we recover with Proposition 5.11 the known result that this class is closed under marginalization (see Proposition 3.4) [14]. For linear SCMs, we have that unique solvability w.r.t. a subset $\mathcal{L}$ holds if and only if ancestral unique solvability w.r.t. $\mathcal{L}$ holds (see Proposition C.4), and hence, a marginalization of a linear SCM always respects the latent projection.

6. Markov properties. In this section, we give a short overview of Markov properties for SCMs with cycles. We make use of the Markov properties that were recently developed by Forré and Mooij [17] for HEDGes, a graphical representation that is similar to the augmented graph of SCMs. We briefly summarize some of their main results and apply them to the class of SCMs. In Appendix A.2, we provide a more thorough introduction and give an intuitive derivation, which can act as an entry point for the reader into the more extensive discussion of Markov properties provided in [17].

Markov properties associate a set of conditional independence relations to a graph. The directed global Markov property for directed acyclic graphs (see Definitions A.4 and A.6), also known as the d-separation criterion [47], is one of the most widely used. It directly extends to a similar property for acyclic directed mixed graphs (ADMGs) [54]. It does not hold in general for cyclic SCMs, however, as was already observed earlier [66, 67].

EXAMPLE 6.1 (Directed global Markov property does not hold for cyclic SCM). One can check that for every solution $\boldsymbol{X}$ of the SCM $\mathcal{M}$ of Example 3.5, X_1 is not independent of X_2 given $\{X_3, X_4\}$. However, the variables X_1 and X_2 are d-separated given $\{X_3, X_4\}$ in $\mathcal{G}(\mathcal{M})$ (see Figure 3). Hence the global directed Markov property does not hold here.

Although some progress has been made in the case of discrete [17, 46, 49] and linear models [17, 24, 28, 59, 65–67], only recently a general directed global Markov property has been introduced for more general cyclic models [17], that is based on σ-separation (see Definitions A.16 and A.20), an extension of d-separation. This notion of σ-separation was derived from the notion of d-separation in the acyclification of the graph [17] (see Definition A.13). The acyclification of a graph generalizes the idea of the collapsed graph developed by Spirtes [66] and can, in particular, be applied to the graphs of SCMs. The main idea of the acyclification is that under the condition that the SCM is uniquely solvable w.r.t. each strongly connected component, we can replace the causal mechanisms of these strongly connected components by their measurable solution functions, which results in an acyclic SCM. This acyclified SCM (see Definition A.11) is observationally equivalent to the original SCM (see Proposition A.12).

EXAMPLE 6.2 (Construction of an observationally equivalent acyclic SCM). The SCM $\mathcal{M}$ of Example 3.5 is uniquely solvable w.r.t. all its strongly connected components, that is, the subsets $\{1\}$, $\{2\}$ and $\{3, 4\}$. Replacing the causal mechanisms of these strongly connected components by their measurable solution functions gives the observationally equivalent SCM $\tilde{\mathcal{M}}$ of Example 4.2. Because $\tilde{\mathcal{M}}$ is acyclic (see Figure 3) we can apply the directed global Markov property to $\tilde{\mathcal{M}}$. The fact that X_1 and X_2 are not d-separated given $\{X_3, X_4\}$ in $\mathcal{G}(\tilde{\mathcal{M}})$ is in line with X_1 being dependent of X_2 given $\{X_3, X_4\}$ for every solution $\boldsymbol{X}$ of $\tilde{\mathcal{M}}$ (and hence of $\mathcal{M}$).

This acyclification preserves solutions, and d-separation in the acyclification can directly be translated into σ-separation on the original graph (see Proposition A.19). This leads to the general directed global Markov property. The following theorem summarizes the main results of [17] applied to SCMs.

THEOREM 6.3 (Global Markov properties for SCMs [17]). *Let $\mathcal{M}$ be a uniquely solvable SCM. Then its observational distribution $\mathbb{P}^{\boldsymbol{X}}$ exists, is unique and the following two statements hold*:

1. $\mathbb{P}^{\boldsymbol{X}}$ *satisfies the* directed global Markov property (*"d-separation criterion"*) *relative to* $\mathcal{G}(\mathcal{M})$ (*see Definition* A.6) *if $\mathcal{M}$ satisfies at least one of the following conditions*:
   (a) *$\mathcal{M}$ is acyclic*;
   (b) *all endogenous spaces $\mathcal{X}_i$ are discrete and $\mathcal{M}$ is ancestrally uniquely solvable*;
   (c) *$\mathcal{M}$ is linear* (*see Definition* C.1), *each of its causal mechanisms $\{f_i\}_{i\in\mathcal{I}}$ has a nontrivial dependence on at least one exogenous variable*, *and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{\mathcal{J}}$.*
2. $\mathbb{P}^{\boldsymbol{X}}$ *satisfies the* general directed global Markov property (*"σ-separation criterion"*) *relative to* $\mathcal{G}(\mathcal{M})$ (*see Definition* A.20) *if $\mathcal{M}$ is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$.*[14]

The general directed global Markov property is generally weaker than the directed global Markov property, since σ-separation implies d-separation. The acyclic case is well known and was first shown in the context of linear-Gaussian structural equation models [29, 70]. The discrete case fixes the erroneous theorem by Pearl and Dechter [49], for which a counterexample was found by Neal [46], by adding the ancestral unique solvability condition, and extends it to allow for bidirected edges in the graph. The linear case is an extension of existing results for the linear-Gaussian setting without bidirected edges [28, 66, 67] to a linear (possibly non-Gaussian) setting with bidirected edges in the graph.

In constraint-based approaches to causal discovery, one usually assumes the converse of the (general) directed global Markov property to hold [48, 68], which is called σ-faithfulness respectively d-faithfulness (see Definitions A.9 and A.23). Meek [38] showed that for multinomial and linear-Gaussian DAG (i.e., acyclic and causally sufficient SCMs) models, d-faithfulness holds for all parameter values up to a measure zero set. Up to our knowledge no such results have been shown in more general parametric or nonparametric settings (neither for d-faitfhulness in acyclic or cyclic settings, nor for σ-faithfulness).

7. Causal interpretation of the graph of SCMs. In Example 4.4, we already saw that sometimes no information in the observational, interventional and even the counterfactual distributions suffices to decide whether a directed path or bidirected edge is present in the graph, or not. Here, we do not attempt to provide a complete characterization of the conditions under which the presence or absence of a directed path or bidirected edge in the graph can be

[14]Since [17] also provides results under the weaker condition that an SCM is solvable (not necessarily uniquely) w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$, one might believe that Theorem 6.3.(2) could be generalized to stating that in that case, any of its observational distributions satisfies the general directed global Markov property. However, that is not true: consider, for example, the SCM $\mathcal{M} = \langle \mathbf{2}, \varnothing, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}}\rangle$ with $f_1(\boldsymbol{x}) = x_1$ and $f_2(\boldsymbol{x}) = x_2$. Then $\mathcal{M}$ is solvable w.r.t. each of its strongly connected components $\{1\}$ and $\{2\}$. The solution with $X_1 = X_2$, where X_2 has a nondegenerate distribution, shows a dependence between X_1 and X_2, and thus $X_1 \perp\!\!\!\perp X_2$ does not hold. In general, all strongly connected components that admit multiple solutions may be dependent on any other variable(s) in the model.

identified from the observational and interventional distributions. Instead, we give sufficient conditions to detect a directed path and bidirected edge in the graph.

In general, cyclic SCMs may have none, one or multiple induced observational distributions, and this may change after intervening in the system. Here, we restrict ourselves to graphs of SCMs where the induced (marginal) observational and interventional distributions are uniquely defined.

7.1. *Directed paths and edges.* For cyclic SCMs, the causal interpretation of the SCM is not always consistent with its graph. This can be illustrated with the SCM $\mathcal{M}$ of Example 5.10. Here, one sees a difference in the marginal distribution $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{1\},\xi_1)}}$ on $\mathcal{X}_4$ for different values of ξ_1, although variable 1 is not an ancestor of variable 4 and each marginal distribution $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{1\},\xi_1)}}$ on $\mathcal{X}_4$ is uniquely defined. This counterintuitive behavior that an intervention on a nonancestor of a variable can change the distribution of that variable was already observed by Neal [46]. However, under a specific unique solvability condition, we obtain a direct causal interpretation for the absence of a directed edge or directed path in the graph of an SCM.

PROPOSITION 7.1 (Sufficient condition for detecting a directed edge in the latent projection of the graph of an SCM). *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$, a subset $\mathcal{O} \subseteq \mathcal{I}$ and $i, j \in \mathcal{O}$ such that $i \neq j$. Let $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, where $I := \mathcal{O} \setminus \{i, j\}$, such that $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is uniquely solvable w.r.t.* $\mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})\setminus i}(j)$. *If there exist values $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ such that both $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}$ and $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\tilde{\xi}_i)}$ induce unique marginal distributions on $\mathcal{X}_j$, and these two induced distributions do not coincide, that is, there exists a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that*

$$\mathbb{P}_{(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}}(X_j \in \mathcal{B}_j) \neq \mathbb{P}_{(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\tilde{\xi}_i)}}(X_j \in \mathcal{B}_j),$$

the directed edge $i \to j$ is present in the latent projection $\mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$ *of $\mathcal{G}(\mathcal{M})$ on $\mathcal{O}$.*

Two cases are of special interest: $\mathcal{O} = \mathcal{I}$, which corresponds with a directed edge $i \to j$ in $\mathcal{G}(\mathcal{M})$, and $\mathcal{O} = \{i, j\}$, which corresponds with a directed path $i \to \cdots \to j$ in $\mathcal{G}(\mathcal{M})$.

The condition in Proposition 7.1 is a sufficient condition for determining whether a directed edge or path is present in the graph. In general, not all directed edges and paths can be identified from the interventional distributions with this sufficient condition. For example, no interventional distribution satisfies the condition of Proposition 7.1 for the SCM $\bar{\mathcal{M}}$ in Example 4.4, although there is a directed edge $1 \to 2$ in the graph $\mathcal{G}(\bar{\mathcal{M}})$.

7.2. *Bidirected edges.* It is well known that there exists a similar sufficient condition for detecting bidirected edges in the graph of an acyclic SCM also known as the common-cause principle (see, e.g., [48]). In the two variables case, this criterion informally states that there exists a bidirected edge between the variables i and j in the graph of the SCM, if the marginal interventional distribution of X_j under the intervention $\mathrm{do}(\{i\}, x_i)$ differs from the conditional distribution of X_j given $X_i = x_i$ (see Example D.13). The following proposition provides a generalization of this sufficient condition for detecting bidirected edges in graphs of SCMs that may include cycles.

PROPOSITION 7.2 (Sufficient condition for detecting a bidirected edge in the latent projection of the graph of an SCM). *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$, a subset $\mathcal{O} \subseteq \mathcal{I}$ and $i, j \in \mathcal{O}$ such that $i \neq j$. Let $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, where $I := \mathcal{O} \setminus \{i, j\}$, such that $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is uniquely solvable w.r.t. both* $\mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ *and* $\mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})\setminus i}(j)$. *Assume* *that for every $\xi_i \in \mathcal{X}_i$ both $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ and $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}$ induce a unique marginal distribution on $\mathcal{X}_j \times \mathcal{X}_i$ and $\mathcal{X}_j$, respectively. If $j \notin \mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ and there exists a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that for every version of the regular conditional probability $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B}_j \mid X_i = \xi_i)$, there exists a value $\xi_i \in \mathcal{X}_i$ such that*

$$\mathbb{P}_{(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}}(X_j \in \mathcal{B}_j) \neq \mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B}_j \mid X_i = \xi_i),$$

then there exists a bidirected edge $i \leftrightarrow j$ in the latent projection $\mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$ *of $\mathcal{G}(\mathcal{M})$ on $\mathcal{O}$.*

This proposition gives a sufficient condition for determining that a bidirected edge is present in the graph. In general, not all bidirected edges in the graph can be identified from the observational, interventional and even the counterfactual distributions, as we saw in Example D.10. In this example, there exists a bidirected edge $1 \leftrightarrow 2 \in \mathcal{G}(\mathcal{M})$ while the density $p(x_2 \mid \mathrm{do}(X_1 = x_1)) = p(x_2 \mid X_1 = x_1)$ for all $x_1 \in \mathcal{X}_1$. For the acyclic setting, the above criterion is generally considered as a universal way to detect a confounder (note that then one can also deal with the case $j \in \mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ by swapping the roles of i and j). If i and j are part of a cycle, the above sufficient condition cannot be applied, and in that case, to the best of our knowledge, no simple sufficient conditions for detecting the presence of a bidirected edge are known.

8. Simple SCMs. In this section, we introduce the well-behaved class of simple SCMs. Simple SCMs satisfy all the local unique solvability conditions to ensure that this class is closed under both perfect intervention and marginalization. They extend the subclass of acyclic SCMs to the cyclic setting, while preserving many of their convenient properties.

DEFINITION 8.1 (Simple SCM). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$ *simple* if it is uniquely solvable w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$.

Loosely speaking, an SCM is simple if any subset of its structural equations can be solved uniquely for its associated variables in terms of the other variables that appear in these equations. An example of a simple SCM is given in Example D.1.

On simple SCMs one can perform any number of marginalizations (see Definition 5.3) in any order (see Proposition 5.4). All these marginalizations respect the latent projection (see Proposition 5.11) and each resulting marginal SCM is again simple. Moreover, we show that this class is closed under intervention and the twin operation.

PROPOSITION 8.2. *The class of simple SCMs is closed under marginalization, perfect intervention and the twin operation.*

The class of simple SCMs contains the acyclic SCMs as a subclass (see Proposition 3.4). In particular, a simple SCM has no self-cycles (see Proposition 3.7), since a self-cycle denotes that that variable cannot be uniquely (up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set) determined by its parents.

From Proposition 8.2, it follows that the results summarized in Theorem 6.3 also apply to all the observational, interventional and counterfactual distributions of simple SCMs.

COROLLARY 8.3 (Global Markov properties for simple SCMs). *Let $\mathcal{M}$ be a simple SCM. Then the*:

1. *observational distribution*,
2. *interventional distribution after perfect intervention on $I \subset \mathcal{I}$*,
3. *counterfactual distribution after perfect intervention on $\tilde{I} \subseteq \mathcal{I} \cup \mathcal{I}'$*,

all exist, are unique and satisfy the general directed global Markov property relative to $\mathcal{G}(\mathcal{M})$, $\mathrm{do}(I)(\mathcal{G}(\mathcal{M}))$ *and* $\mathrm{do}(\tilde{I})(\mathrm{twin}(\mathcal{G}(\mathcal{M})))$, *respectively. Moreover, if* $\mathcal{M}$ *satisfies at least one of the three conditions* (1*a*), (1*b*), (1*c*) *of Theorem* 6.3, *then they also obey the directed global Markov property relative to* $\mathcal{G}(\mathcal{M})$, $\mathrm{do}(I)(\mathcal{G}(\mathcal{M}))$ *and* $\mathrm{do}(\tilde{I})(\mathrm{twin}(\mathcal{G}(\mathcal{M})))$, *respectively.*

Many of these properties are also shown to hold for the class of *modular SCMs* [17], which contains, in particular, the class of simple SCMs (see Appendix A.3 for more details).

Moreover, simple SCMs satisfy the unique solvability conditions of Propositions 7.1 and 7.2, which allows us to define the causal relationships for simple SCMs in terms of its graph.

DEFINITION 8.4 (Causal relationships for simple SCMs). Let $\mathcal{M}$ be a simple SCM.

1. If there exists a directed edge $i \to j \in \mathcal{G}(\mathcal{M})$, that is, $i \in \mathrm{pa}(j)$, then we call i *a direct cause of* j *according to* $\mathcal{M}$;
2. If there exists a directed path $i \to \cdots \to j$ in $\mathcal{G}(\mathcal{M})$, that is, $i \in \mathrm{an}(j)$, then we call i *a cause of* j *according to* $\mathcal{M}$;
3. If there exists a bidirected edge $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$, then we call i *and* j *(latently) confounded according to* $\mathcal{M}$.

In summary, we have the following sufficient conditions for determining the different causal and confoundedness relationships according to a specific simple SCM $\mathcal{M}$.

COROLLARY 8.5 (Sufficient conditions for the presence of causal and confoundedness relationships for simple SCMs). *Let* $\mathcal{M}$ *be a simple SCM and* $i, j \in \mathcal{I}$ *such that* $i \neq j$ *and* $I := \mathcal{I} \setminus \{i, j\}$. *Then*:

1. *If there exist values* $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ *and* $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ *and a measurable set* $\mathcal{B}_j \subseteq \mathcal{X}_j$ *such that*
$$\mathbb{P}_{(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}}(X_j \in \mathcal{B}_j) \neq \mathbb{P}_{(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\tilde{\xi}_i)}}(X_j \in \mathcal{B}_j),$$
then i *is a direct cause of* j *according to* $\mathcal{M}$, *that is,* $i \to j \in \mathcal{G}(\mathcal{M})$;
2. *If there exist values* $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ *and a measurable set* $\mathcal{B}_j \subseteq \mathcal{X}_j$ *such that*
$$\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{i\},\xi_i)}}(X_j \in \mathcal{B}_j) \neq \mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{i\},\tilde{\xi}_i)}}(X_j \in \mathcal{B}_j),$$
then i *is a cause of* j *according to* $\mathcal{M}$, *that is,* $i \to \cdots \to j$ *in* $\mathcal{G}(\mathcal{M})$;
3. *If* $j \notin \mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ *and there exist a value* $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ *and a measurable set* $\mathcal{B}_j \subseteq \mathcal{X}_j$ *such that for every version of the regular conditional probability* $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B}_j \mid X_i = \xi_i)$ *there exists a value* $\xi_i \in \mathcal{X}_i$ *such that*
$$\mathbb{P}_{(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}}(X_j \in \mathcal{B}_j) \neq \mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B}_j \mid X_i = \xi_i),$$
then i *and* j *are confounded according to* $\mathcal{M}$, *that is,* $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$.

For simple SCMs, it is in general not possible to identify all the causal and confoundedness relationships in the graph from the observational, interventional or even the counterfactual distributions. Examples 4.4 and D.10 show that this is already impossible for acyclic SCMs without further assumptions.

Finally, there is a connection between SCMs and potential outcomes [63] that generalizes to the cyclic setting. One of the consequences of Proposition 8.2 is that all counterfactuals are defined for a simple SCM (even if it is cyclic). This allows us to define potential outcomes in terms of a simple SCM in the following way.

DEFINITION 8.6 (Potential outcome). Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be a simple SCM, $I \subseteq \mathcal{I}$ a subset, $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value and $\boldsymbol{E}$ a random variable such that $\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$. The *potential outcome under the perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$ is defined as $\boldsymbol{X}_{\boldsymbol{\xi}_I} := \boldsymbol{g}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(\boldsymbol{E}_{\mathrm{pa}(\mathcal{I})})$, where $\boldsymbol{g}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ is a measurable solution function for $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$.

9. Discussion. In this paper, we studied the basic properties of SCMs in the presence of cycles and latent variables without restricting to linear functional relationships between the variables. We saw that cyclic SCMs behave differently in many aspects than acyclic SCMs. Indeed, in the presence of cycles, many of the convenient properties of acyclic SCMs do not hold in general: SCMs do not always have a solution; they do not always induce unique observational, interventional and counterfactual distributions; a marginalization does not always exist, and if it exists the marginal model does not always respect the latent projection; they do not always satisfy a Markov property and their graphs are not always consistent with their causal semantics.

We introduced various notions of (unique) solvability and showed that under appropriate (unique) solvability conditions, many of the operations and results for the acyclic setting can be extended to SCMs with cycles. For example, we introduced several equivalence relations between SCMs to compare SCMs at different levels of abstraction, we showed how to define marginal SCMs on a subset of the variables that are (in various ways) equivalent to the original SCM, we discussed under which conditions the distributions satisfy the (general) directed global Markov property relative to their graphs and we showed under which conditions the graph of an SCM can be interpreted causally. Most of these results are shown under sufficient conditions that are not necessary (e.g., for the marginalization operation this was shown in Example D.11). It may therefore be possible to further relax some of the conditions.

These insights led us to introduce the more well-behaved class of simple SCMs, which forms an extension of the class of acyclic SCMs to the cyclic setting that preserves many of its convenient properties: simple SCMs induce unique observational, interventional and counterfactual distributions; the class of simple SCMs is closed under both perfect intervention and marginalization; the marginalization respects the latent projection; the induced distributions obey the general directed global Markov property and obey the directed global Markov property in the acyclic, discrete and linear case. This class does not contain SCMs that have self-cycles and graphs of simple SCMs have a direct and intuitive causal interpretation.

One key property of simple SCMs is that the solutions always satisfy the conditional independencies implied by σ-separation. By simply replacing d-separation with σ-separation, it turns out that one can directly extend results and algorithms for acyclic SCMs to the more general class of simple SCMs. For example, adjustment criteria (including the back-door criterion), Pearl's do-calculus and Tian's ID algorithm for the identification of causal effects have been extended recently to the class of modular SCMs, which contains the class of simple SCMs [19]. Several causal discovery algorithms have already been proposed that work with simple SCMs, for example, the first constraint-based causal discovery algorithm that can deal with cycles and nonlinear functional relationships [18]. Also, Local Causal Discovery (LCD) [9], Y-structures [35] and the Joint Causal Inference framework (JCI) all apply to simple SCMs [44] even though they were originally developed for acyclic SCMs only. Recently, it has been shown that even the well-known Fast Causal Inference (FCI) algorithm [69, 75] is directly applicable to simple SCMs [41] and provides a consistent estimate of the Markov equivalence class (under the faithfulness assumption). Moreover, a method for constructing nonlinear simple SCMs using neural networks and sampling from them has been proposed [18]. This illustrates that the class of simple SCMs forms a convenient and practical extension of the class of acyclic SCMs that can be used for the purposes of causal modeling, reasoning, discovery and prediction.

We hope that this work will provide the foundations for a general theory of statistical causal modeling with SCMs. Future work might consist of reparametrizing and reducing the space of the exogenous variables of an SCM while preserving the causal and counterfactual semantics; extending and generalizing the identifiability results for (direct) causes and confounders; extending the graphs of SCMs to represent selection bias; proving completeness results for some Markov properties for a subclass of SCMs that contains cycles.

Acknowledgments. The authors are grateful to Bernhard Schölkopf and Robin Evans for stimulating discussions, and to Noud de Kroon, Tineke Blom and Alexander Ly for providing helpful comments on earlier drafts. We thank two anonymous reviewers and the Associate Editor for helpful comments.

Funding. S. Bongers and J.M. Mooij are supported in part by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410 and VENI grant 639.031.036). P. Forré and J.M. Mooij are supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^{o} 639466). J. Peters is supported by research grants from VILLUM FONDEN (18968) and the Carlsberg Foundation.

SUPPLEMENTARY MATERIAL

Supplement to "Foundations of structural causal models with cycles and latent variables" (DOI: 10.1214/21-AOS2064SUPP; .pdf). This Supplementary Material contains a summary of the basic terminology and results for causal graphical models, additional (unique) solvability properties, some results for linear SCMs, other examples, the proofs of all theoretical results and the measurable selection theorems that are used in several proofs.

REFERENCES

[1] BALKE, A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence* (*AAAI*-94) **1** 230–237. AAAI Press, Menlo Park.

[2] BECKERS, S. and HALPERN, J. Y. (2019). Abstracting causal models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (*AAAI*-19) **33** 2678–2685. AAAI Press, Menlo Park.

[3] BLOM, T., BONGERS, S. and MOOIJ, J. M. (2019). Beyond structural causal models: Causal constraints models. In *Proceedings of the* 35*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-19) (R. P. Adams and V. Gogate, eds.). AUAI Press.

[4] BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0996025 https://doi.org/10.1002/9781118619179

[5] BONGERS, S., BLOM, T. and MOOIJ, J. M. (2021). Causal modeling of dynamical systems. Preprint. Available at arXiv:1803.08784v3 [cs.AI].

[6] BONGERS, S., FORRÉ, P., PETERS, J. and MOOIJ, J. M. (2021). Supplement to "Foundations of structural causal models with cycles and latent variables." https://doi.org/10.1214/21-AOS2064SUPP

[7] BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann*. *Statist*. **42** 2526–2556. MR3277670 https://doi.org/10.1214/14-AOS1260

[8] BYRNE, R. M. J. (2007). *The Rational Imagination*: *How People Create Alternatives to Reality*. *A Bradford Book*. MIT Press, Cambridge, MA.

[9] COOPER, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Min*. *Knowl*. *Discov*. **1** 203–224.

[10] DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *Int*. *Stat*. *Rev*. **70** 161–189.

[11] DUNCAN, O. D. (1975). *Introduction to Structural Equation Models*: *Studies in Population*. Academic Press, New York. MR0398558

[12] EATON, D. and MURPHY, K. (2007). Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (M. Meila and X. Shen, eds.). *Proceedings of Machine Learning Research* **2** 107–114.

[13] EBERHARDT, F., HOYER, P. and SCHEINES, R. (2010). Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.). *Proceedings of Machine Learning Research* **9** 185–192.

[14] EVANS, R. J. (2016). Graphs for margins of Bayesian networks. *Scand*. *J*. *Stat*. **43** 625–648. MR3543314 https://doi.org/10.1111/sjos.12194

[15] EVANS, R. J. (2018). Margins of discrete Bayesian networks. *Ann*. *Statist*. **46** 2623–2656. MR3851750 https://doi.org/10.1214/17-AOS1631

[16] FISHER, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica* **38** 73–92.

[17] FORRÉ, P. and MOOIJ, J. M. (2017). Markov properties for graphical models with cycles and latent variables. Preprint. Available at arXiv:1710.08775 [math.ST].

[18] FORRÉ, P. and MOOIJ, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the* 34*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-18) (A. Globerson and R. Silva, eds.). AUAI Press.

[19] FORRÉ, P. and MOOIJ, J. M. (2019). Causal calculus in the presence of cycles, latent confounders and selection bias. In *Proceedings of the* 35*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-19) (R. P. Adams and V. Gogate, eds.). AUAI Press.

[20] FOYGEL, R., DRAISMA, J. and DRTON, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *Ann*. *Statist*. **40** 1682–1713. MR3015040 https://doi.org/10.1214/12-AOS1012

[21] GOLDBERGER, A. S. and DUNCAN, O. D. (1973). *Structural Equation Models in the Social Sciences*. Seminar Press, New York.

[22] HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. MR0007954 https://doi.org/10.2307/1905714

[23] HALPERN, J. (1998). Axiomatizing causal reasoning. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (*UAI*-98) (G. Cooper and S. Moral, eds.) 202–210. Morgan Kaufmann, San Francisco, CA, USA.

[24] HYTTINEN, A., EBERHARDT, F. and HOYER, P. O. (2012). Learning linear cyclic causal models with latent variables. *J*. *Mach*. *Learn*. *Res*. **13** 3387–3439. MR3005891

[25] HYTTINEN, A., HOYER, P. O., EBERHARDT, F. and JÄRVISALO, M. (2013). Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* (*UAI*-13) (A. Nicholson and P. Smyth, eds.) 301–310. AUAI Press, Corvallis, OR, USA.

[26] IWASAKI, Y. and SIMON, H. A. (1994). Causality and model abstraction. *Artificial Intelligence* **67** 143–194. MR1281657 https://doi.org/10.1016/0004-3702(94)90014-0

[27] KECHRIS, A. S. (1995). *Classical Descriptive Set Theory*. *Graduate Texts in Mathematics* **156**. Springer, New York. MR1321597 https://doi.org/10.1007/978-1-4612-4190-4

[28] KOSTER, J. T. A. (1996). Markov properties of nonrecursive causal models. *Ann*. *Statist*. **24** 2148–2177. MR1421166 https://doi.org/10.1214/aos/1069362315

[29] KOSTER, J. T. A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand*. *J*. *Stat*. **26** 413–431. MR1712043 https://doi.org/10.1111/1467-9469.00157

[30] LACERDA, G., SPIRTES, P. L., RAMSEY, J. and HOYER, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (*UAI*-08) (D. McAllester and P. Myllymaki, eds.) 366–374. AUAI Press, Corvallis, OR, USA.

[31] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Clarendon Press, Oxford. MR1419991

[32] LAURITZEN, S. L., DAWID, A. P., LARSEN, B. N. and LEIMER, H. G. (1990). Independence properties of directed Markov fields. *Networks* **20** 491–505. MR1064735 https://doi.org/10.1002/net.3230200503

[33] LEWIS, D. K. (1979). Counterfactual dependence and time's arrow. *Noûs* **13** 455–476. https://doi.org/10.2307/2215339

[34] MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann*. *Statist*. **37** 3133–3164. MR2549555 https://doi.org/10.1214/09-AOS685

[35] MANI, S. (2006). A Bayesian local causal discovery framework. PhD thesis, Univ. Pittsburg.

[36] MASON, S. J. (1953). Feedback theory—Some properties of signal flow graphs. In *Proceedings of the IRE* **41** 1144–1156. IEEE.

[37] MASON, S. J. (1956). Feedback theory—Further properties of signal flow graphs. In *Proceedings of the IRE* **44** 920–926. IEEE.

[38] MEEK, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (*UAI*-95) (P. Besnard and S. Hanks, eds.) 411–418. Morgan Kaufmann, San Francisco, CA, USA.

[39] MOGENSEN, S. W. and HANSEN, N. R. (2020). Markov equivalence of marginalized local independence graphs. *Ann. Statist.* **48** 539–559. MR4065173 https://doi.org/10.1214/19-AOS1821

[40] MOGENSEN, S. W., MALINSKY, D. and HANSEN, N. R. (2018). Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence* (*UAI*-18) (A. Globerson and R. Silva, eds.) AUAI Press.

[41] MOOIJ, J. M. and CLAASSEN, T. (2020). Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Proceedings of the* 36*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-20) (J. Peters and D. Sontag, eds.) **124** 1159–1168. PMLR.

[42] MOOIJ, J. M. and HESKES, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the* 29*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-13) (A. Nicholson and P. Smyth, eds.) 431–439. AUAI Press, Corvallis, OR, USA.

[43] MOOIJ, J. M., JANZING, D. and SCHÖLKOPF, B. (2013). From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the* 29*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-13) (A. Nicholson and P. Smyth, eds.) 440–448. AUAI Press.

[44] MOOIJ, J. M., MAGLIACANE, S. and CLAASSEN, T. (2020). Joint causal inference from multiple contexts. *J. Mach. Learn. Res.* **21** Paper No. 99, 108. MR4119167

[45] MOOIJ, J. M., PETERS, J., JANZING, D., ZSCHEISCHLER, J. and SCHÖLKOPF, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.* **17** Paper No. 32, 102. MR3491126

[46] NEAL, R. M. (2000). On deducing conditional independence from d-separation in causal graphs with feedback. *J. Artificial Intelligence Res.* **12** 87–91. MR1745047 https://doi.org/10.1613/jair.689

[47] PEARL, J. (1985). A constraint propagation approach to probabilistic reasoning. In *Proceedings of the First Conference on Uncertainty in Artificial Intelligence* (*UAI*-85) (L. Kanal and J. Lemmer, eds.) 31–42. AUAI Press, Corvallis, OR, USA.

[48] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

[49] PEARL, J. and DECHTER, R. (1996). Identifying independence in causal graphs with feedback. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence* (*UAI*-96) (E. Horvitz and F. Jensen, eds.) 420–426. Morgan Kaufmann, San Francisco, CA, USA.

[50] PEARL, J. and MACKENZIE, D. (2018). *The Book of Why*: *The New Science of Cause and Effect*. Basic Books, New York. MR3791446

[51] PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of Causal Inference*: *Foundations and Learning Algorithms*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3822088

[52] PETERS, J., MOOIJ, J. M., JANZING, D. and SCHÖLKOPF, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15** 2009–2053. MR3231600

[53] PFISTER, N., BAUER, S. and PETERS, J. (2019). Learning stable and predictive structures in kinetic systems. *Proc. Natl. Acad. Sci. USA* **116** 25405–25411. MR4047351 https://doi.org/10.1073/pnas.1905688116

[54] RICHARDSON, T. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157. MR1963898 https://doi.org/10.1111/1467-9469.00323

[55] RICHARDSON, T. and SPIRTES, P. (1999). Automated discovery of linear feedback models. In *Computation*, *Causation*, *and Discovery* (C. Glymour and G. F. Cooper, eds.) 253–302. AAAI Press, Menlo Park, CA. MR1696455

[56] RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166 https://doi.org/10.1214/aos/1031689015

[57] RICHARDSON, T. S. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence* (*UAI*-96) (E. Horvitz and F. Jensen, eds.) 454–461. Morgan Kaufmann, San Francisco, CA, USA. MR1617227

[58] RICHARDSON, T. S. (1996). Discovering cyclic causal structure. Technical Report No. CMU-PHIL-68, Carnegie Mellon Univ.

[59] RICHARDSON, T. S. (1996). Models of feedback: Interpretation and discovery. Ph.D. thesis, Carnegie Mellon Univ.

[60] RICHARDSON, T. S. and ROBINS, J. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report No. 128, Center for Statistics and the Social Sciences.

[61] ROESE, N. J. (1997). Counterfactual thinking. *Psychol. Bull.* **121** 133–148.

[62] RUBENSTEIN, P. K., WEICHWALD, S., BONGERS, S., MOOIJ, J. M., JANZING, D., GROSSE-WENTRUP, M. and SCHÖLKOPF, B. (2017). Causal consistency of structural equation models. In *Proceedings of the* 33*rd Conference on Uncertainty in Artificial Intelligence* (*UAI*-17) (G. Elidan and K. Kersting, eds.). AUAI Press.

[63] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

[64] SHPITSER, I. and PEARL, J. (2008). Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* **9** 1941–1979. MR2447308

[65] SPIRTES, P. (1993). Directed cyclic graphs, conditional independence, and non-recursive linear structural equation models. Technical Report No. CMU-PHIL-35, Carnegie Mellon Univ.

[66] SPIRTES, P. (1994). Conditional independence in directed cyclic graphical models for feedback. Technical Report No. CMU-PHIL-54, Carnegie Mellon Univ.

[67] SPIRTES, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (*UAI*-95) (P. Besnard and S. Hanks, eds.) 499–506. Morgan Kaufmann, San Francisco, CA, USA.

[68] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

[69] SPIRTES, P., MEEK, C. and RICHARDSON, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery* (C. Glymour and G. F. Cooper, eds.) 211–252. AAAI Press, Menlo Park, CA. MR1689972

[70] SPIRTES, P., RICHARDSON, T., MEEK, C., SCHEINES, R. and GLYMOUR, C. (1998). Using path diagrams as a structural equation modelling tool. *Sociol. Methods Res.* **27** 182–225.

[71] TIAN, J. (2002). Studies in causal reasoning and learning. Technical Report No. R-309, Cognitive Systems Laboratory, Univ. California, Los Angeles, USA.

[72] TIAN, J. and PEARL, J. (2001). Causal discovery from changes. In *Proceedings of the* 17*th Conference in Uncertainty in Artificial Intelligence* (*UAI*-01) (J. Breese and D. Koller, eds.) 512–521. Morgan Kaufmann, San Francisco, CA, USA.

[73] VERMA, T. S. (1993). Graphical aspects of causal models. Technical Report No. R-191. Computer Science Department, Univ. California, Los Angeles, USA.

[74] WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.

[75] ZHANG, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* **172** 1873–1896. MR2459793 https://doi.org/10.1016/j.artint.2008.08.001