FOUNDATIONS OF STRUCTURAL CAUSAL MODELS WITH CYCLES AND LATENT VARIABLES

By Stephan Bongers[*,§], Patrick Forré[†,§], Jonas Peters[‡,¶], Bernhard Schölkopf[‖] and Joris M. Mooij[*,†,§]

University of Amsterdam[§]*, University of Copenhagen*[¶] *and Max Planck Institute for Intelligent Systems*[‖]

Structural causal models (SCMs), also known as (non-parametric) 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 (non-parametric) structural equation models (SEMs), are widely used for causal modeling purposes (Bollen, 1989; Spirtes, Glymour and Scheines, 2000; Pearl, 2009; Peters, Janzing and Schölkopf, 2017). They form the basis for many statistical methods that aim at inferring knowledge of the underlying causal structure from data (see e.g., Maathuis et al., 2009; Mooij and Heskes, 2013; Peters et al., 2014; Bühlmann, Peters and Ernest, 2014; Mooij et al., 2016). 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 (Wright, 1921), econometrics (Haavelmo, 1943), electrical engineering (Mason, 1953, 1956), and the social sciences (Goldberger and Duncan, 1973; Duncan, 1975).

Acyclic SCMs, also known as *recursive SEMs*, form a special well-studied class of SCMs that gen-

[*]Supported in part by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410 and VENI grant 639.031.036).

[†]Supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 639466).

[‡]Supported by a research grant (18968) from VILLUM FONDEN.

eralize causal Bayesian networks (Pearl, 2009) to allow for modeling latent confounders. They have many convenient properties (see e.g., Pearl, 1985; Lauritzen et al., 1990; Verma, 1993; Lauritzen, 1996; Richardson, 2003; Evans, 2016, 2018): (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, feedback loops between observed variables are present (see e.g., Haavelmo, 1943; Mason, 1953, 1956; Mooij and Heskes, 2013; Pfister, Bauer and Peters, 2019). For those systems, SCMs with cycles (or "non-recursive SEMs") may form an appropriate model class (Mooij, Janzing and Schölkopf, 2013; Bongers and Mooij, 2018). In contrast to their acyclic counterparts, SCMs with cycles have enjoyed less attention in the literature and are not as well understood, since in general none of the above properties (i)-(vi) holds in this class. However, some progress has been made in the case of discrete (Pearl and Dechter, 1996; Neal, 2000) and linear models (Spirtes, 1993, 1994, 1995; Koster, 1996; Hyttinen, Eberhardt and Hoyer, 2012), and more recently, their Markov properties have been elucidated in more generality (Forré and Mooij, 2017).

Contributions. The purpose of this paper is to formally describe and study the class of SCMs in the presence of cycles and latent variables while allowing for non-linear functional relationships between those variables. We investigate to which extent and under which sufficient conditions each of the properties (i)-(vi) still holds in this class. Our aim is to provide the foundations for a general theory of statistical causal modeling with SCMs. In the next paragraphs, we describe our contributions in more detail.

When cyclic functional relationships hold between variables, one encounters various technical complications. 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, which leads to ambiguity (Halpern, 1998). 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. Even worse, it is not clear whether the solutions of the structural equations of an SCM are measurable if cycles are present. One of our contributions consists in providing sufficient 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 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 by Balke and Pearl (1994). These different equivalence relations formalize the different levels of abstraction in the so-called "causal hierarchy" (Shpitser and Pearl, 2008; Pearl and Mackenzie, 2018).

An important elementary operation in probability theory is "marginalization": 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. One can show that the induced observational and interventional distributions of the marginalized SCM coincide with the marginals of the distributions induced by the original SCM. In other words, the operation of marginalization preserves the probabilistic and causal semantics (restricted to the remaining variables). While this was already known for acyclic SCMs (see Verma, 1993; Spirtes et al., 1998; Evans, 2016, 2018, a.o.), we show here that for cyclic SCMs this is not true without further assumptions. In (Forré and Mooij, 2017) it is shown that under the strong assumption of "modularity" a marginalization can also be defined for cyclic SCMs and that it preserves the probabilistic and causal semantics. We show here that marginalizations of SCMs can be defined under weaker assumptions. Intuitively, the idea is to think of the subset of endogenous variables that are going to be marginalized out as a subsystem that interacts with the rest of the system. Under a certain unique solvability condition, one can ignore the internals of this subsystem and treat it effectively as a "black box", which has a unique output for every possible input, and this can be substituted into the structural equations of the remaining variables. We show that the original and the marginal SCM 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" (Verma, 1993; Tian, 2002; Evans, 2016). In general, the marginalization of an SCM does not respect the latent projection of its associated graph, but we show that it does under certain ancestral unique solvability conditions.

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 (Lauritzen, 1996), one prominent example being the "d-separation criterion", also known as the directed global Markov property, which was originally derived for Bayesian networks (Pearl, 1985). 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 (1994). His solution in terms of "collapsed graphs" was recently generalized and reformulated for a general class of causal graphical models (Forré and Mooij, 2017) 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 (Pearl, 2009). As was already observed by Neal (2000), the causal interpretation of SCMs with cycles can be counter-intuitive, as the causal semantics

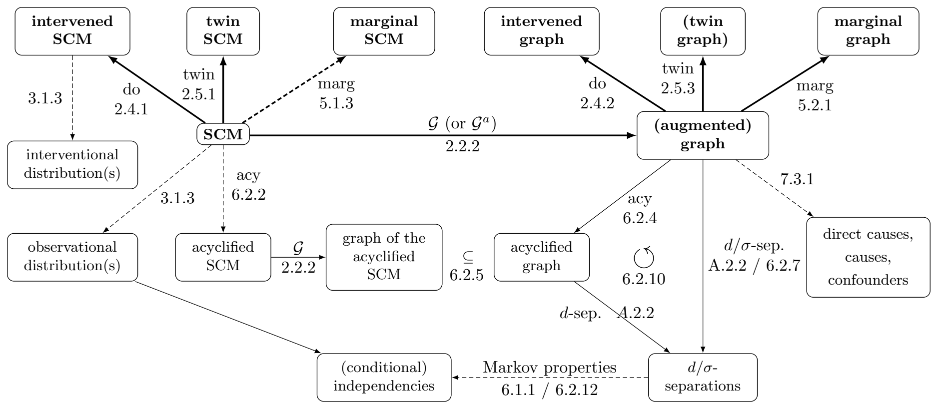

Fig 1: *Overview of the objects constructed from an SCM and the mappings between them. The numbers correspond to the definition, proposition or theorem of the corresponding object, mapping, or result. When an arrow is dashed, the relation only holds under non-trivial assumptions that can be found in the corresponding definition or theorem. The symbol "$\subseteq$" stands for the subgraph of a directed mixed graph (see Definition A.1.1 in the Appendix A) and the symbol "$\circlearrowleft$" 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.*

under interventions no longer needs to be compatible with the structure imposed by the functional relations between the variables. We resolve this here by showing that under certain ancestral unique solvability conditions the causal interpretation of SCMs is consistent with its graph.

Overall, we conclude that 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. It preserves much of the simplicity of the theory for the acyclic setting, while allowing for cycles. 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.

The scope of this paper is limited to establishing the foundations for statistical causal modeling with cyclic SCMs. For a detailed discussion of causal reasoning, causal discovery and causal prediction with cyclic SCMs we refer the reader to other literature (e.g., Richardson, 1996a,b; Richardson and Spirtes, 1999; Eberhardt, Hoyer and Scheines, 2010; Hyttinen, Eberhardt and Hoyer, 2012; Hyttinen et al., 2013; Foygel, Draisma and Drton, 2012). Several recent results (generalizations of the do-calculus, adjustment criteria and an identification algorithm) for modular SCMs (Forré and Mooij, 2018, 2019) directly apply to the subclass of simple SCMs as well. Finally, it is interesting to note that many causal discovery algorithms that have been designed for the acyclic case also apply to simple SCMs with no or only minor changes (Mooij, Magliacane and Claassen, 2019).

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.

SCMs	do	twin	marg
$\mathcal{G}, \mathcal{G}^a$	Prop. 2.4.3	Prop. 2.5.4	(Prop. 5.2.6)
do	Prop. 2.4.4.(1)	Prop. 2.5.2	Prop. 5.1.5
twin	$\cdots$	-	Prop. 5.1.6
marg	$\cdots$	$\cdots$	Prop. 5.1.4

Graphs	do	twin	marg
do	Prop. 2.4.4.(1)	Prop. 2.5.5	Prop. 5.2.3
twin	$\cdots$	-	Prop. 5.2.4
marg	$\cdots$	$\cdots$	Prop. 5.2.2

Table 1: *Overview of the commutativity results of different pairs of mappings, defined on SCMs (top table) and on graphs (bottom table). 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 5.2.6 (in parentheses) is not a commutativity result but a weaker relation. The graphical twin operator is only defined for directed graphs. All results apply under the assumptions stated in the corresponding proposition.*

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, how it relates to self-cycles and define the class of simple SCMs. 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. Section 7 discusses the causal interpretation of the graphs of SCMs and Section 8 concludes with a brief discussion.

The appendices provide an introduction to the basic terminology for directed graphical models in Appendix A; some results for linear SCMs in Appendix B; the proofs of all the theoretical results in Appendix C; and some lemmas and measurable selection theorems that are used in several proofs in Appendix D.

2. Structural causal models. In this section, we provide the basic definitions and properties of structural causal models (SCMs). We start in Section 2.1 by formally defining SCMs and their solutions, and introducing a natural equivalence relation on SCMs. In Section 2.2, we introduce graphical representations of an SCM. We then show in Section 2.3 that each SCM has a representative with a sparsest graph, its structurally minimal representation. In Section 2.4 we introduce perfect interventions, a key notion that grounds the causal semantics of SCMs. We finish in Section 2.5 with a definition of the twin SCM and how this gives rise to counterfactuals.

2.1. *Definition of a structural causal model and its solutions.* Our definition of SCMs slightly deviates from existing definitions (Bollen, 1989; Pearl, 2009; Spirtes, Glymour and Scheines, 2000) because we make the definition of the SCM independent of the random variables that solve it. This will enable us to deal with the various technical complications that arise in the presence of cycles.

Definition 2.1.1 (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 codomains of the endogenous variables, where each codomain $\mathcal{X}_i$ is a standard measurable space,*[2]
4. *$\boldsymbol{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ is the product of the codomains of the exogenous variables, where each codomain $\mathcal{E}_j$ 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 mechanisms*,*
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 $j \in \mathcal{J}$.*[3]

Although it is common to assume the absence of cyclic functional relations (see Definition 2.2.4), we make no such assumption here. When allowing for cycles, it turns out that the most natural setting is obtained when allowing for self-cycles as well. We will discuss self-cycles in more detail in Section 3.2.

In structural causal models, the functional relationships between variables are expressed in terms of (deterministic) equations involving the causal mechanisms.

Definition 2.1.2 (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*

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

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

Readers already familiar with structural causal models will note that we are still missing an important ingredient: the random variables that express solutions of SCMs.

Definition 2.1.3 (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. *$\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$, i.e., the distribution of $\boldsymbol{E}$ is equal to $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$,*[4] *and*
2. *the* structural equations *are satisfied, i.e.,*

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

[1]We often use boldface for variables that have multiple components, e.g., vectors or tuples in a Cartesian product.

[2]A standard measurable space is a measurable space $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ that is isomorphic to a measurable space $(\boldsymbol{\mathcal{Y}}, \mathcal{B}(\boldsymbol{\mathcal{Y}}))$, where $\boldsymbol{\mathcal{Y}}$ is a Polish space (i.e., a separable completely metrizable space) and $\mathcal{B}(\boldsymbol{\mathcal{Y}})$ are the Borel subsets of $\boldsymbol{\mathcal{Y}}$ (i.e., the σ-algebra generated by the open sets in $\boldsymbol{\mathcal{Y}}$). Throughout this paper, when we say that $\boldsymbol{\mathcal{X}}$ is a standard measurable space, then we implicitly assume that there exists a σ-algebra $\boldsymbol{\Sigma}$ such that $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ is a standard measurable space. In several proofs we assume without loss of generality that the standard measurable space *is* a Polish space $\boldsymbol{\mathcal{Y}}$ with σ-algebra $\mathcal{B}(\boldsymbol{\mathcal{Y}})$. Examples of standard measurable spaces are the open and closed subsets of $\mathbb{R}^d$, and the finite sets with the usual complete metric. See Appendix D for more details.

[3]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 $\mathbb{P}_{\mathbf{1}}$.

[4]Here the components E_j of $\boldsymbol{E}$ are mutually independent, since $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ is a product measure on $\prod_{j \in \mathcal{J}} \mathcal{E}_j$.

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* $\boldsymbol{X}$. Note that 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 shows.

Example 2.1.4 (Cyclic SCMs). *For brevity, we use throughout this paper the notation* $\boldsymbol{n} := \{1, 2, \dots, n\}$ *for* $n \in \mathbb{N}$. *Let* $\mathcal{M} = \langle \boldsymbol{2}, \boldsymbol{1}, \mathbb{R}^2, \mathbb{R}, f, \mathbb{P}_{\mathbb{R}} \rangle$ *be an SCM*[5] *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* X *any arbitrary random variable 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. This ambiguity is illustrated by the following example.

Example 2.1.5 (Equivalent SCMs have the same solutions). *Let* $\mathcal{M} = \langle \boldsymbol{1}, \boldsymbol{1}, \mathcal{X}, \mathcal{E}, f, \mathbb{P}_{\mathcal{E}} \rangle$ *be the SCM with* $\mathcal{X} = \mathcal{E} = \{-1, 0, 1\}$, $\mathbb{P}_{\mathcal{E}}(\{-1\}) = \mathbb{P}_{\mathcal{E}}(\{1\}) = \frac{1}{2}$ *and* $f(x, e) = e^2 + e - 1$. *Take for* $\tilde{\mathcal{M}}$ *the SCM* $\mathcal{M}$ *but with a different causal mechanism* $\tilde{f}(e) = e$. *Then the set of solutions of the structural equations agree for both SCMs for* $e \in \{-1, +1\}$, *while they differ only for* $e = 0$, *which occurs with probability zero. Hence, a pair of random variables* (X, E) *is a solution of* $\mathcal{M}$ *if and only if it is of* $\tilde{\mathcal{M}}$.

To accommodate for this it seems natural not to differentiate between causal mechanisms that have different solutions on at most a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set of exogenous variables. This leads to a natural equivalence relation between the causal mechanisms.

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, $\boldsymbol{\mathcal{X}}_{\emptyset}$ and $\boldsymbol{\mathcal{E}}_{\emptyset}$ are defined by the singleton $\{1\}$. Moreover, for 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.

Definition 2.1.6 (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* $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *are* equivalent, *denoted by* $\mathcal{M} \equiv \tilde{\mathcal{M}}$, *if for all* $i \in I$, *for* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-almost every* $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$

[5] In our examples, we will abuse notation by using non-disjoint 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.

and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

Note that two equivalent SCMs can only differ in terms of their causal mechanism. Importantly, equivalent SCMs have the same solutions.[6]

This equivalence relation $\equiv$ on the set of all SCMs gives rise to the quotient set of equivalence classes of SCMs. In this paper we prove several properties and define several operations and relations on the quotient set of equivalence classes of SCMs. A common approach for proving a certain property for an equivalence class of SCMs is that we start by proving that the property holds for a representative of the equivalence class, and then show that it holds for any other element of that equivalence class. Similarly, in order to define a certain operation (or relation) on the equivalence class of SCMs, we usually start by defining the operation on an SCM and then show that this operation preserves the equivalence relation.

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 (Wright, 1921). The structural properties of the functional relations between variables modeled by an SCM are described 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 2.2.1. *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* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*, we have*

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

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.

DEFINITION 2.2.2 (Graph and augmented graph). *Given an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$*, 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* $v \in \mathcal{I}$*;*[8]

[6]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 non-unique or even non-existent. 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.

[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]By Definition 2.2.1, $u \in \mathcal{I} \cup \mathcal{J}$ cannot be a parent of $v \in \mathcal{J}$.

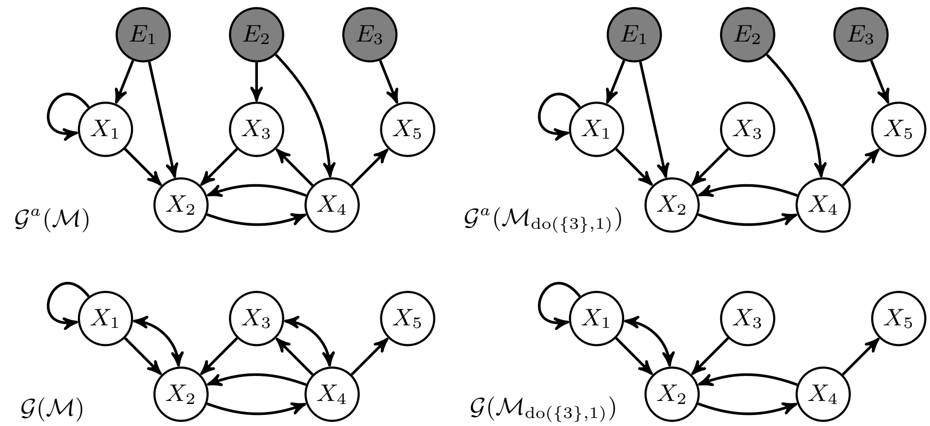

Fig 2: *The augmented graph (top) and the graph (bottom) of the SCM $\mathcal{M}$ of Example 2.2.3 (left) and of the intervened SCM $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ of Example 2.4.5 (right).*

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 $v \in \mathcal{I}$.*

In particular, the augmented graph contains no directed edges between the exogenous variables, because they are not functionally related through the causal mechanism. These definitions map $\mathcal{M}$ to $\mathcal{G}^a(\mathcal{M})$ and $\mathcal{G}(\mathcal{M})$. We call these mappings the *augmented graph mapping* $\mathcal{G}^a$ and the *graph mapping* $\mathcal{G}$ respectively. 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 2.2.3 (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*

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

where $\alpha \in \mathbb{R}$ and $\mathbb{P}_{\mathbb{R}^3}$ a product of non-degenerate[9] *probability measures over $\mathbb{R}^3$. The augmented graph $\mathcal{G}^a(\mathcal{M})$ and the graph $\mathcal{G}(\mathcal{M})$ of $\mathcal{M}$ are depicted*[10] *in Figure 2 (left). Observe that if the product probability measure $\mathbb{P}_{\mathbb{R}^3}$ had been a degenerate measure, then the directed edges from the exogenous variables to the endogenous variables in the augmented graph $\mathcal{G}^a(\mathcal{M})$ and the bidirected edges in the graph $\mathcal{G}(\mathcal{M})$ would have disappeared.*

As is illustrated in this example, the augmented graph is 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 often write the sets $\mathrm{pa}_{\mathcal{G}^a(\mathcal{M})}(\mathcal{U})$, $\mathrm{ch}_{\mathcal{G}^a(\mathcal{M})}(\mathcal{U})$,

[9] A probability measure $\mathbb{P}$ over a measurable space $\boldsymbol{\mathcal{X}}$ is called *degenerate* if it is the Dirac measure $\boldsymbol{\delta_x}$ for some point $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, which is defined on the measurable set $\boldsymbol{\mathcal{U}} \subseteq \boldsymbol{\mathcal{X}}$ by $\boldsymbol{\delta_x}(\boldsymbol{\mathcal{U}}) := \mathbf{1}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{x})$ with $\mathbf{1}_{\boldsymbol{\mathcal{U}}}$ the indicator function.

[10] For visualizing an (augmented) graph, we stick to 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 will still use the random variables notation in the (augmented) graph in the case that the SCM has no solution at all.

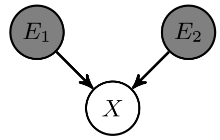

Fig 3: *The augmented graph of the acyclic SCM $\mathcal{M}$ of Example 2.3.2.*

$\mathrm{an}_{\mathcal{G}^a(\mathcal{M})}(\mathcal{U})$, etc., for some subset $\mathcal{U} \subseteq \mathcal{I} \cup \mathcal{J}$, simply as $\mathrm{pa}(\mathcal{U})$, $\mathrm{ch}(\mathcal{U})$, $\mathrm{an}(\mathcal{U})$ for brevity when there is no danger of confusion.

Definition 2.2.4. *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) (Richardson, 2003).

Most of the existing literature focuses on acyclic SCMs (see Pearl, 2009; Spirtes, Glymour and Scheines, 2000; Verma, 1993; Tian, 2002; Evans, 2016, 2018, a.o.). In the structural equation model (SEM) literature, acyclic SCMs are referred to as *recursive* SEMs and cyclic SCMs as *non-recursive* SEMs (Bollen, 1989). Acyclic SCMs are also known as *semi-Markovian* SCMs (Pearl, 2009; Tian, 2002). 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. For these models it was first shown that they satisfy several Markov properties (Pearl, 2009; Lauritzen et al., 1990; Tian, 2002). Acyclic SCMs have considerable technical advantages over SCMs with cycles. For example, they always induce a unique observational distribution, the class of acyclic SCMs is closed under intervention and marginalization, and they satisfy several equivalent Markov properties (Pearl, 1985, 2009; Richardson, 2003; Evans, 2016). These properties make acyclic SCMs a convenient class to work with, which may explain the focus on acyclic SCMs in most of the literature.

2.3. *The structurally minimal representation.* In this section we show that there always exists a representative of the equivalence class of an SCM for which each component of the causal mechanism does not depend on its non-parents (see also Peters, Janzing and Schölkopf, 2017).

Definition 2.3.1 (Structurally minimal SCM). *Consider the SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$. 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 $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$.*

The next example illustrates that not all SCMs are structurally minimal.

Example 2.3.2 (A non-structurally minimal SCM). *Consider an SCM $\mathcal{M} = \langle \mathbf{1}, \mathbf{2}, \mathbb{R}, \mathbb{R}^2, f, \mathbb{P}_{\mathbb{R}^2} \rangle$ with causal mechanism $f(x, \boldsymbol{e}) = -x + e_1 + e_2$ and $\mathbb{P}_{\mathbb{R}^2}$ a product of non-degenerate probability measures over $\mathbb{R}$. Its augmented graph $\mathcal{G}^a(\mathcal{M})$ is depicted in Figure 3. The variable x is not a parent of itself, while $f(x, \boldsymbol{e})$ depends on x, and hence $\mathcal{M}$ is not structurally minimal. However, the equivalent SCM $\tilde{\mathcal{M}}$ defined by replacing the causal mechanism of $\mathcal{M}$ by $\tilde{f}(x, \boldsymbol{e}) = \frac{1}{2}(e_1 + e_2)$ is structurally minimal.*

In general, there always exists an equivalent structurally minimal SCM like in the previous example.

PROPOSITION 2.3.3 (Existence of a 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. Then there exists an 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 and with $\tilde{\mathcal{M}} \equiv \mathcal{M}$.*

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 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 (2009) that we refer to as "perfect" interventions. Other types of interventions, like mechanism changes (Tian and Pearl, 2001), fat-hand interventions (Eaton and Murphy, 2007), activity interventions (Mooij and Heskes, 2013), and stochastic versions of all these are at least as relevant, but we will not consider those here.

DEFINITION 2.4.1 (Perfect intervention on an SCM). *Given 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 $I \subseteq \mathcal{I}$ of endogenous variables and a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, the* perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *maps $\mathcal{M}$ to the* intervened 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 defined by*

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

This operation $\mathrm{do}(I, \boldsymbol{\xi}_I)$ preserves the equivalence relation (see Definition 2.1.6) on the set of all SCMs and hence this mapping induces a well-defined mapping on the set of equivalence classes of SCMs. In contrast to the work of (Rubenstein et al., 2017; Beckers and Halpern, 2019; Blom, Bongers and Mooij, 2019), who assume that it is only possible to intervene on a specific subset of endogenous variables, here 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 2.4.2 (Perfect intervention on a directed mixed graph). *Given a directed mixed graph*[11] *$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ and a subset $I \subseteq \mathcal{V}$, the perfect intervention* $\mathrm{do}(I)$ *maps $\mathcal{G}$ to the* intervened 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 $\tilde{\mathcal{B}} := \mathcal{B} \setminus \{v \leftrightarrow i : v \in \mathcal{V}, i \in \mathcal{I}\}$.*

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

PROPOSITION 2.4.3. *Given an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and a subset $I \subseteq \mathcal{I}$ of endogenous variables and a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, then $\big(\mathcal{G}^a \circ \mathrm{do}(I, \boldsymbol{\xi}_I)\big)(\mathcal{M}) = \big(\mathrm{do}(I) \circ \mathcal{G}^a\big)(\mathcal{M})$ and $\big(\mathcal{G} \circ \mathrm{do}(I, \boldsymbol{\xi}_I)\big)(\mathcal{M}) = \big(\mathrm{do}(I) \circ \mathcal{G}\big)(\mathcal{M})$.*

[11]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.1.1 in the Appendix A for a more precise definition).

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

PROPOSITION 2.4.4. *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 give an SCM with a solution after intervention.

EXAMPLE 2.4.5 (Intervened SCM and its graphs). *Consider the SCM $\mathcal{M}$ of Example 2.2.3 which has a solution if and only if $\alpha > 0$. 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*

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

for which the augmented graph $\mathcal{G}^a(\mathcal{M}_{\mathrm{do}(\{3\},1)})$ and the graph $\mathcal{G}(\mathcal{M}_{\mathrm{do}(\{3\},1)})$ are 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. The reverse is also possible: doing another perfect intervention $\mathrm{do}(\{4\},1)$ *on $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ gives again an SCM with a solution for $\alpha > 0$.*

Remember that for each solution $\boldsymbol{X}$ of an SCM $\mathcal{M}$ we called 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.[12] For example, the SCM $\mathcal{M}$ of Example 2.2.3 has two different observational distributions if $\alpha > 0$. Similarly, an intervened SCM may induce a distribution that is 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 $\boldsymbol{X}$*.[13]

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*. This twin SCM was first introduced in the "twin network" method by Balke and Pearl (1994).

DEFINITION 2.5.1 (Twin SCM). *Given an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, 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 defined by $\tilde{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{e}) := \big(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{e}), \boldsymbol{f}(\boldsymbol{x}',\boldsymbol{e})\big)$.

[12] 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 for example Definition 7.1.1 in Pearl (2009)). One has to be careful: although these SCMs induce a unique observational distribution, they generally do not induce a unique distribution after a perfect intervention.

[13] In the literature, one often finds the notation $p(\boldsymbol{x})$ and $p(\boldsymbol{x}\,|\,\mathrm{do}(\boldsymbol{X}_I = \boldsymbol{x}_I))$ for the densities of the observational and interventional distribution, respectively, in case these are uniquely defined by the SCM (e.g. Pearl, 2009).

By definition, the twin operation on SCMs preserves the equivalence relation $\equiv$ on SCMs. The perfect intervention and the twin operation for SCMs commute with each other in the following way.

PROPOSITION 2.5.2. *For an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and a subset $I \subseteq \mathcal{I}$ of endogenous variables and a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, we have the identity $(\mathrm{do}(I \cup I', \boldsymbol{\xi}_{I \cup I'}) \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin} \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M})$, where I' is the copy of I in $\mathcal{I}'$ and $\boldsymbol{\xi}_{I'} = \boldsymbol{\xi}_I$.*

Whenever the intervened twin SCM $(\mathcal{M}^{\mathrm{twin}})_{\mathrm{do}(\tilde{I}, \boldsymbol{\xi}_{\tilde{I}})}$, 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* $\mathrm{do}(\tilde{I}, \boldsymbol{\xi}_{\tilde{I}})$ *associated to* $(\boldsymbol{X}, \boldsymbol{X}')$.

Next, we define a 'twin' operation that operates on directed graphs.

DEFINITION 2.5.3 (Twin graph). *Given a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a subset $I \subseteq \mathcal{V}$ such that $J := \mathcal{V} \setminus I$ is exogenous, i.e., $\mathrm{pa}_{\mathcal{G}}(J) = \emptyset$, we define the* twin graph w.r.t. *I as the directed graph* $\mathrm{twin}(I)(\mathcal{G}) = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$, *where*

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

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

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

These two twin operations are compatible with the augmented graph mapping.

PROPOSITION 2.5.4. *Given an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, then $(\mathcal{G}^a \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin}(\mathcal{I}) \circ \mathcal{G}^a)(\mathcal{M})$.*

We have the following result analogous to Proposition 2.5.2.

PROPOSITION 2.5.5. *Given a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and subsets $I \subseteq K \subseteq \mathcal{V}$ such that $\mathrm{pa}_{\mathcal{G}}(\mathcal{V} \setminus K) = \emptyset$, then $(\mathrm{do}(I \cup I') \circ \mathrm{twin}(K))(\mathcal{G}) = (\mathrm{twin}(K) \circ \mathrm{do}(I))(\mathcal{G})$, where I' is the copy of I in K'.*

Both twin operations preserve acyclicity.

PROPOSITION 2.5.6. *For SCMs and directed graphs we have that acyclicity is preserved under the twin operation.*

A typical counterfactual query is a question of the form "What is $p(\boldsymbol{X}_{M'} = \boldsymbol{x}_{M'} \,|\, \mathrm{do}(\boldsymbol{X}_{I'} = \boldsymbol{\xi}_{I'}), \mathrm{do}(\boldsymbol{X}_J = \boldsymbol{\eta}_J), \boldsymbol{X}_{K'} = \boldsymbol{x}_{K'}, \boldsymbol{X}_L = \boldsymbol{x}_L)$?", where $J, L \subseteq \mathcal{I}$ and $I', K', M' \subseteq \mathcal{I}'$ are all disjoint and we assume that there exists a unique interventional distribution for which a density exists. Such a query reads "Given that in the actual world we performed a perfect intervention $\mathrm{do}(\boldsymbol{X}_J = \boldsymbol{\eta}_J)$ and afterwards observed $\boldsymbol{X}_L = \boldsymbol{x}_L$, what would have been the probability of $\boldsymbol{X}_{M'} = \boldsymbol{x}_{M'}$ in the counterfactual world in which instead we had done the perfect intervention $\mathrm{do}(\boldsymbol{X}_{I'} = \boldsymbol{\xi}_{I'})$ and had afterwards observed $\boldsymbol{X}_{K'} = \boldsymbol{x}_{K'}$?".

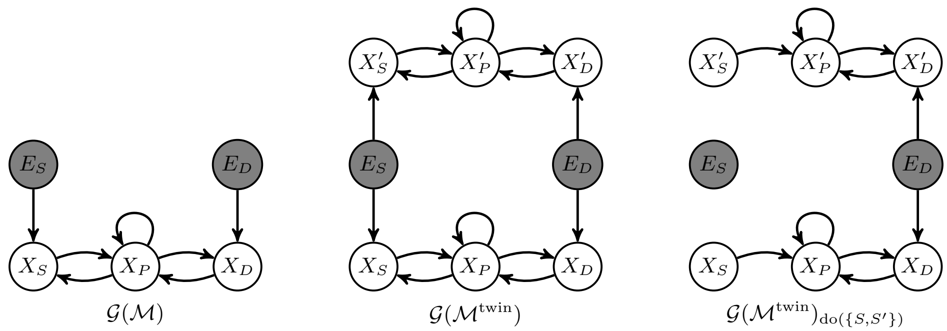

Fig 4: *The augmented graph of the SCM $\mathcal{M}$ (left), its twin SCM $\mathcal{M}^{\text{twin}}$ (center) and the intervened twin SCM $(\mathcal{M}^{\text{twin}})_{\text{do}(S,s),\text{do}(S',s')}$ (right) of Example 2.5.7.*

An example is provided by the following well-known market equilibrium model from economics, which has been thoroughly discussed in the literature (see e.g., Richardson and Robins, 2014). This illustrates how self-cycles enrich the class of SCMs, and that counterfactuals can be sensibly formulated even in cyclic SCMs with self-loops.

EXAMPLE 2.5.7 (Price, supply and demand). *Consider a simple model of price, supply and demand of a quantity of a product, specified by the SCM $\mathcal{M} = \langle \{P, S, D\}, \{S, D\}, \mathbb{R}^3, \mathbb{R}^2, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with causal mechanism defined by*

$$\begin{aligned} f_D(\boldsymbol{x}, \boldsymbol{e}) &:= \alpha_D + \beta_D x_P + e_D \\ f_S(\boldsymbol{x}, \boldsymbol{e}) &:= \alpha_S + \beta_S x_P + e_S \\ f_P(\boldsymbol{x}, \boldsymbol{e}) &:= x_P + (x_D - x_S)\,. \end{aligned}$$

Here x_D denotes the demand and x_S the supply of a quantity of a product, while the price of the product is denoted by x_P. The demanded and supplied quantities are determined by the price, and price is determined implicitly by the condition that demanded and supplied quantities should be equal (a situation known as "market equilibrium"). Exogenous influences on demand and supply are modeled by e_D and e_S respectively, $\beta_D < 0$ is the reciprocal of the slope of the demand curve, and $\beta_S > 0$ is the reciprocal of the slope of the supply curve. Note how we use a self-cycle for P in order to implement the equilibrium equation $X_D = X_S$ as the causal mechanism for the price P.[14]

Note that $\mathcal{M}$ is uniquely solvable. As an example of a counterfactual query, consider

$$\mathbb{P}(X'_P \,|\, \text{do}(X'_S = s'), \text{do}(X_S = s), X_P = p)\,,$$

*which denotes the conditional distribution of X'_P given $X_P = p$ of a solution of $\mathcal{M}^{\text{twin}}_{\text{do}(\{S',S\},(s',s))}$. In words: how would—*ceteris paribus*—price have been distributed, had we intervened to set supplied quantities equal to s', given that actually we intervened to set supplied quantities equal to s and observed that this led to price p? A straightforward calculation shows that this counterfactual distribution of price is the Dirac measure on $x'_P = p + (s' - s)/\beta_D$.*

The augmented graphs of the SCM, its twin graph, and its intervened twin graph are depicted in Figure 4.

[14] Richardson and Robins (2014) argue that this market equilibrium model cannot be modeled as an SCM. We observe that it can, as long as one allows for self-cycles.

The interpretation of counterfactual statements has received a lot of attention in the literature (Lewis, 1979; Roese, 1997; Byrne, 2007; Balke and Pearl, 1994; Pearl, 2009). One of the reasons is that there exist SCMs that induce the same observational and interventional distributions, but differ in their counterfactual statements (Dawid, 2002) (see also Example 4.3.5). This raises the question how one would ever be able to estimate the parameters of such an SCM from data. An alternative graphical approach to counterfactuals is the framework of Single World Intervention Graphs (SWIGs) (Richardson and Robins, 2013). This can be generalized to the cyclic setting as well, but we will not do so here.

3. Solvability.

In this section we describe the notions of solvability and unique solvability with respect to a subset of the endogenous variables of an SCM. These notions 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. Moreover, we discuss several properties of (unique) solvability and we introduce the class of simple SCMs, which have very convenient solvability properties.

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 $\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}$. A solution function w.r.t. $\mathcal{O}$ assigns each input value $(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})})$ of this subsystem to a specific output value $\boldsymbol{x}_{\mathcal{O}}$ of the subsystem. This is formalized as follows.

Definition 3.1.1 (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 $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*

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

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.

Example 3.1.2 (Different cases of solvability). *Consider the SCM $\mathcal{M}$ of Example 2.2.3 and the subset of endogenous variables $\{2,3,4\}$ which is depicted by the box around the nodes in the augmented graph in Figure 5 (left). For each input value $x_1 \in \mathcal{X}_1$ and $(e_1, e_2) \in \boldsymbol{\mathcal{E}}_{\{1,2\}}$ of the box, the structural equations for the variables $\{2,3,4\}$ have a unique output for x_2, x_3 and x_4, which is given by the mapping $\boldsymbol{g}_{\{2,3,4\}} : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $\boldsymbol{g}_{\{2,3,4\}}(x_1, e_1, e_2) := (x_1 + e_1 + e_2, -x_1 - e_1 - e_2, x_1 + e_1 + 2e_2)$. The existence of such a (clearly measurable) mapping means that $\mathcal{M}$ is solvable w.r.t. $\{2,3,4\}$. Solvability does not require the uniqueness on the output variables, for example if we consider the subset $\{1\}$ and take $\alpha > 0$, then there exist two measurable solution functions $g_1^+, g_1^- : \mathbb{R}^1 \to \mathbb{R}^1$ of $\mathcal{M}$ w.r.t. $\{1\}$ defined by $g_1^{\pm}(e_1) := \pm\sqrt{\alpha^{-1} e_1^2}$. In general, solvability does not*

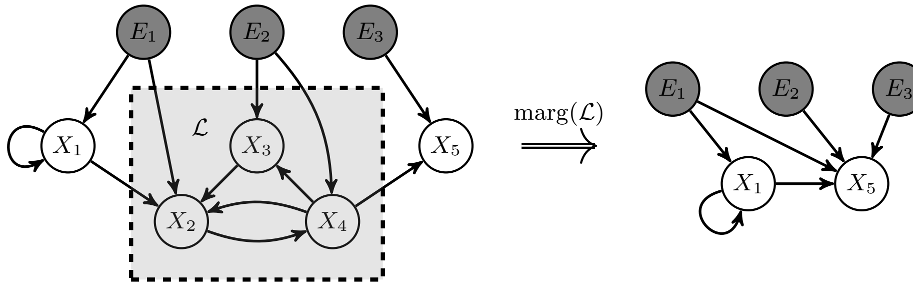

Fig 5: *The augmented graphs of the SCM $\mathcal{M}$ (left) and $\tilde{\mathcal{M}}$ (right) of Example 2.2.3, 3.1.2 and 5.1.2, where the SCM $\tilde{\mathcal{M}}$ is a marginalization of $\mathcal{M}$ w.r.t. $\mathcal{L}$.*

hold w.r.t. every subset. For example, $\mathcal{M}$ is not solvable w.r.t. the subset $\{2,4\}$, because the equality $x_1 + x_3 + e_1 + e_2 = 0$ does not hold for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$.

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

THEOREM 3.1.3 (Sufficient and necessary conditions for solvability). *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$. Then the following are equivalent:*

1. *$\mathcal{M}$ has a solution (see Definition 2.1.3);*
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. *there exists a measurable mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*
$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{e}) \quad \Longrightarrow \quad \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) \,.$$
4. *$\mathcal{M}$ is solvable (see Definition 3.1.1).*

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 $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ 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 D.8 in the Appendix D).

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 $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{g}(\boldsymbol{E}_{\mathrm{pa}(\mathcal{I})})$ 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 $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{E}_{\mathrm{pa}(\mathcal{I})})$ holds a.s., e.g. if $\boldsymbol{X}$ is a non-trivial mixture of such solutions, as in the following example.

EXAMPLE 3.1.4 (Mixtures of solutions are solutions). *Consider an SCM $\mathcal{M} = \langle \mathbf{1}, \emptyset, \mathbb{R}, \mathbf{1}, f, \mathbb{P}_{\mathbf{1}} \rangle$ with causal mechanism $f : \mathcal{X} \times \mathcal{E} \to \mathcal{X}$ defined by $f(x,e) = x - x^2 + 1$. There exist only two measurable solution functions $g_{\pm} : \mathcal{E} \to \mathcal{X}$ for $\mathcal{M}$, defined by $g_{\pm}(e) = \pm 1$. Let $X : \Omega \to \mathbb{R}$ be a random variable that is a non-trivial mixture of point masses on $\{-1, +1\}$. Then X is a solution of $\mathcal{M}$, however neither $g_+(E) = X$ a.s., nor $g_-(E) = X$ a.s., for any random variable E such that $\mathbb{P}^E = \mathbb{P}_{\mathcal{E}}$.*

The condition of solvability w.r.t. a strict subset is in general not a sufficient and necessary condition for the existence of a (global) solution of the SCM. Consider for example the SCM $\mathcal{M}$ in Example 2.2.3 where we take $\alpha = 0$. This SCM is solvable w.r.t. $\{2, 3, 4\}$, however it is not solvable, and hence does not have any solution. However, similarly to the sufficient (and necessary) condition (2) in Theorem 3.1.3 for solvability, we do have a sufficient condition for solvability w.r.t. a (strict) subset.

PROPOSITION 3.1.5 (Sufficient condition for solvability w.r.t. a subset). *Given an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and a subset* $\mathcal{O} \subseteq \mathcal{I}$. *If 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 topological space*

$$\boldsymbol{\mathcal{S}}_{(\boldsymbol{e}, \boldsymbol{x}_{\setminus \mathcal{O}})} := \{\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} : \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})\},$$

with the subspace topology induced by $\boldsymbol{\mathcal{X}}_{\mathcal{O}}$ *is non-empty and* σ*-compact,*[15] *then* $\mathcal{M}$ *is solvable w.r.t.* $\mathcal{O}$.

Again, in the cyclic case the existence of a measurable solution function seems much less straightforward than in the acyclic case.

For many purposes, this condition of σ-compactness suffices since it contains for example all countable discrete spaces, every interval of the real line, and moreover all the Euclidean spaces. It also suffices as a sufficient condition for the marginalization of an SCM (see Theorem 3.2.5 and Definition 5.1.3). For larger solution spaces, we refer the reader to (Kechris, 1995). For the class of linear SCMs (see Definition B.1 in the Appendix B), we do have a sufficient and necessary condition for solvability w.r.t. a (strict) subset of $\mathcal{I}$ (see Proposition B.2).

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 $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ is unique up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set.

DEFINITION 3.2.1 (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* $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ *such that for* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-almost every* $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ *and for all* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

Note that 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. $\mathcal{O}$ as well, 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 play no important role in the theory of acyclic SCMs.

PROPOSITION 3.2.2. *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* $\mathcal{O} \subseteq \mathcal{I}$.

The following example illustrates that also cyclic SCMs can be uniquely solvable w.r.t. a subset.

[15] A topological space $\boldsymbol{\mathcal{X}}$ is called *σ-compact* if it is the union of a countable set of compact topological spaces.

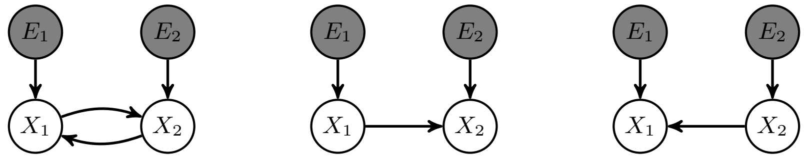

Fig 6: *The augmented graph of the SCM $\mathcal{M}$ (left) of Example 3.2.3, and of the SCMs $\tilde{\mathcal{M}}$ (middle) and $\hat{\mathcal{M}}$ (right) of Example 4.1.2. The SCMs $\mathcal{M}$, $\tilde{\mathcal{M}}$ and $\hat{\mathcal{M}}$ are all observationally equivalent, but not interventionally equivalent.*

EXAMPLE 3.2.3 (Cyclic SCM, uniquely solvable w.r.t. each subset). *Consider the linear SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{2}, \mathbb{R}^2, \mathbb{R}^2, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with causal mechanism given by*

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = \alpha x_2 + e_1 \,, \quad f_2(\boldsymbol{x}, \boldsymbol{e}) = \beta x_1 + e_2 \,,$$

with $\alpha, \beta \neq 0$, $\alpha\beta \neq 1$, and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^{\boldsymbol{E}}$ with normal distributions $E_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $E_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $E_1 \perp\!\!\!\perp E_2$. The augmented graph of $\mathcal{M}$ is depicted in Figure 6 (left). This SCM includes a cycle and is uniquely solvable w.r.t. every subset.

For linear SCMs the unique solvability condition w.r.t. a subset of endogenous variables is equivalent to a matrix invertibility condition (see Proposition B.3 in the Appendix B).

A self-cycle at an endogenous variable in the (augmented) graph of an SCM indicates that the SCM is not uniquely solvable w.r.t. that variable, and vice versa.

PROPOSITION 3.2.4 (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 $i \in \mathcal{I}$ iff $\mathcal{G}^a(\mathcal{M})$ (or $\mathcal{G}(\mathcal{M})$) has no self-cycle $i \to i$ at $i \in \mathcal{I}$.*

In Theorem 3.1.3 we gave sufficient and necessary conditions for (global) solvability. In Proposition 3.1.5 we provided a sufficient condition for solvability w.r.t. a strict subset $\mathcal{O} \subsetneq \mathcal{I}$. The next theorem states that under the additional uniqueness assumption there exists a sufficient and necessary condition for unique solvability w.r.t. any subset, and moreover, that all solutions of a uniquely solvable SCM induce the same observational distribution.

THEOREM 3.2.5 (Sufficient and necessary conditions for unique solvability). *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and a subset of endogenous variables $\mathcal{O} \subseteq \mathcal{I}$. 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.

3.3. *(Unique) solvability properties.* We know that a solvable SCM has a solution (see Theorem 3.1.3), and moreover that all solutions of a uniquely solvable SCM induce the same observational distribution (see Theorem 3.2.5). The property of (unique) solvability, however, is in general is not preserved under perfect intervention. For example, a (uniquely) solvable SCM can lead to a non-uniquely solvable SCM after intervention, which either has no solution or has solutions with multiple different induced distributions.

EXAMPLE 3.3.1 (Solvability is not preserved under perfect intervention). *Consider the SCM* $\mathcal{M} = \langle \mathbf{2}, \emptyset, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *with the following causal mechanism*

$$f_1(\boldsymbol{x}) = x_1 + x_1^2 - x_2 + 1\,, \quad f_2(\boldsymbol{x}) = x_2(1 - \mathbf{1}_{\{0\}}(x_1)) + 1\,.$$

This SCM is (uniquely) solvable. Doing a perfect intervention $\mathrm{do}(\{1\}, \xi_1)$ *for some* $\xi_1 \neq 0$*, however, leads to an intervened model* $\mathcal{M}_{\mathrm{do}(\{1\},\xi_1)}$ *that is not solvable. Performing instead the perfect intervention* $\mathrm{do}(\{2\}, \xi_2)$ *for some* $\xi_2 > 1$ *leads also to a non-uniquely solvable SCM* $\mathcal{M}_{\mathrm{do}(\{2\},\xi_2)}$ *which has two solutions* $(X_1, X_2) = (\pm\sqrt{\xi_2 - 1}, \xi_2)$.

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 non-intervened endogenous variables.

PROPOSITION 3.3.2. *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 let* $\mathcal{O} \subseteq \mathcal{I}$*. If* $\mathcal{M}$ *is (uniquely) solvable w.r.t.* $\mathcal{O}$*, then for every* $\boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}$ *the intervened SCM* $\mathcal{M}_{\mathrm{do}(\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}, \boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}})}$ *is (uniquely) solvable w.r.t.* $\mathrm{pa}(\mathcal{O}) \cup \mathcal{O}$*. Moreover, for every* $\boldsymbol{\xi}_{\mathcal{I}\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\mathcal{O}}$ *the intervened SCM* $\mathcal{M}_{\mathrm{do}(\mathcal{I}\setminus\mathcal{O}, \boldsymbol{\xi}_{\mathcal{I}\setminus\mathcal{O}})}$ *is (uniquely) solvable.*

As we saw in Proposition 3.2.2 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 2.4.4). Moreover, since acyclicity is preserved under the twin operation (see Proposition 2.5.6), an acyclic SCM induces unique observational, interventional and counterfactual distributions.

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 $\mathcal{W} \subsetneq \mathcal{O}$ as can be seen in the following example.

EXAMPLE 3.3.3 (Solvability is not preserved under taking a strict sub- or superset). *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *where the causal mechanism is given by*

$$f_1(\boldsymbol{x}) = x_1 \cdot (1 - \mathbf{1}_{\{1\}}(x_2)) + 1\,,\ f_2(\boldsymbol{x}) = x_2\,,\ f_3(\boldsymbol{x}) = x_3 \cdot (1 - \mathbf{1}_{\{-1\}}(x_2)) + 1\,.$$

This SCM is (uniquely) solvable w.r.t. the subsets $\{1, 2\}$, $\{2, 3\}$*, however it is not (uniquely) solvable w.r.t. the subsets* $\{1\}$, $\{3\}$ *and* $\{1, 2, 3\}$*, and not uniquely solvable w.r.t.* $\{2\}$.

In general, (unique) solvability is not preserved under taking the union and the intersection. Example 3.3.3 gives an example where (unique) solvability is not preserved under taking the union. Even for taking the union of disjoint subsets, (unique) solvability is not preserved (see Example 3.2.3 with $\alpha = \beta = 1$). The next example illustrates that (unique) solvability is in general not preserved under taking the intersection.

EXAMPLE 3.3.4 (Solvability is not preserved under taking the intersection). *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *where the causal mechanism is given by*

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

Then $\mathcal{M}$ *is (uniquely) solvable w.r.t.* $\{1, 2\}$ *and* $\{2, 3\}$*, however it is not (uniquely) solvable w.r.t. their intersection.*

3.4. *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[16] $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$.

DEFINITION 3.4.1 (Ancestral (unique) solvability). *We call an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$* 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 3.4.2 (Solvability is equivalent to ancestral solvability). *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and a subset $\mathcal{O} \subseteq \mathcal{I}$. Then $\mathcal{M}$ is solvable w.r.t. $\mathcal{O}$ iff $\mathcal{M}$ 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 3.4.3 (Unique solvability w.r.t. $\mathcal{O}$ does not imply ancestral unique solvability w.r.t. $\mathcal{O}$). *Consider the SCM $\mathcal{M} = \langle \mathbf{3}, \mathbf{1}, \mathbb{R}^3, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ where the causal mechanism is given by*

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

and $\mathbb{P}_{\mathbb{R}}$ is the standard-normal measure on $\mathbb{R}$. This SCM is uniquely solvable w.r.t. the set $\{1, 2\}$, and thus solvable w.r.t. this set. Although it is solvable w.r.t. the ancestral subset $\{2\}$ in $\mathcal{G}(\mathcal{M})_{\{1,2\}}$ it is not uniquely solvable w.r.t. this subset. Hence, it is not ancestrally uniquely solvable w.r.t. $\{1, 2\}$.

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 B.4 in the Appendix B).

We have seen that unique solvability is not preserved under taking a union. Under an additional assumption, it is.

PROPOSITION 3.4.4 (Combining measurable solution functions on different sets). *Given 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 two ancestral subsets $\mathcal{A}, \tilde{\mathcal{A}} \subseteq \mathcal{O}$ in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$. If $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{A}$, $\tilde{\mathcal{A}}$ and $\mathcal{A} \cap \tilde{\mathcal{A}}$, then $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{A} \cup \tilde{\mathcal{A}}$.*

A consequence of this property is that to check whether an SCM is ancestrally uniquely solvable w.r.t. $\mathcal{O}$, it suffices to check that it is uniquely solvable w.r.t. the ancestral subsets for each node in $\mathcal{O}$.

COROLLARY 3.4.5. *Given an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and a subset $\mathcal{O} \subseteq \mathcal{I}$. Then $\mathcal{M}$ is ancestrally uniquely solvable w.r.t. $\mathcal{O}$ iff $\mathcal{M}$ is uniquely solvable w.r.t. $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i)$ for every $i \in \mathcal{O}$.*

[16]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 $\mathcal{A} \subseteq \mathcal{O}$.

In general, the property of ancestral unique solvability is not preserved under perfect intervention, as is illustrated in the following example.

EXAMPLE 3.4.6 (Ancestral unique solvability is not preserved under perfect intervention). *Consider the SCM $\mathcal{M}$ of Example 3.4.3, but with the causal mechanism component f_2 replaced by $f_2(\boldsymbol{x}, e) = x_3$. This SCM $\mathcal{M}$ is ancestrally uniquely solvable, however $\mathcal{M}_{\mathrm{do}(\{2\},\xi_2)}$ for some $\xi_2 \in \mathbb{R}$ is not ancestrally uniquely solvable.*

The notion of ancestral unique solvability will appear in various results in Sections 5 and 6.

3.5. *Simple SCMs.* We now introduce simple SCMs, which form a convenient subclass of SCMs that may have cycles and confounders. They are convenient because they induce unique observational, interventional and counterfactual distributions, while this class is closed under perfect interventions. Moreover, this class contains the class of acyclic SCMs and is free from SCMs that have self-cycles.

DEFINITION 3.5.1 (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 $\mathcal{O} \subseteq \mathcal{I}$.*

This class of simple SCMs contains the acyclic SCMs as a subclass (see Proposition 3.2.2). Acyclic SCMs have the convenient properties that they are closed under perfect intervention and the twin operation (see Proposition 2.4.4 and Proposition 2.5.6), and they always induce unique observational, interventional and counterfactual distributions (see Proposition 3.2.2 and 3.2.5). These convenient properties also hold for simple SCMs.

PROPOSITION 3.5.2. *Simplicity is preserved under perfect intervention and the twin operation.*

COROLLARY 3.5.3. *All observational, interventional and counterfactual distributions of a simple SCM exist and are unique.*

In particular, this implies that the class of simple SCMs does not contain SCMs that have self-cycles (see Proposition 3.2.4).

We will see later that simple SCMs have the following additional convenient properties: they are closed under marginalization (see Section 5); they obey several Markov properties (see Section 6); and their graphs express their causal semantics (see Section 7). Many of these properties were also shown to hold for the class of *modular SCMs*, which can be seen as an SCM together with an additional structure of a "compatible system of solution functions" (Forré and Mooij, 2017). In the upcoming sections we show that without this additional structure, these properties do not hold in general, and we will investigate to which extent and under which conditions each of these properties still hold. However, for the special case of simple SCMs one can always trivially construct such a compatible system of solution functions, and hence, each simple SCM induces in particular a modular SCM.

As a final note, we would like to point out a connection between SCMs and potential outcomes that generalizes to the cyclic setting. The potential outcome framework (Rubin, 1974) is often used in the causal inference literature. One of the consequences of Corollary 3.5.3 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 3.5.4 (Potential outcome). *Given a simple SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, *a subset* $I \subseteq \mathcal{I}$, *a value* $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ *and a random variable* $\boldsymbol{E}$ *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* $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$.

4. Equivalences. In Section 2 we already encountered an equivalence relation on the class of SCMs (see Definition 2.1.6). The (augmented) graph of an SCM is preserved under this equivalence relation as well as its solutions and its induced observational, interventional and counterfactual distributions. 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 4.1.1 (Observational equivalence). *Consider two SCMs* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and* $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$. $\mathcal{M}$ *and* $\tilde{\mathcal{M}}$ *are* observationally equivalent with respect to $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, *denoted by* $\mathcal{M} \equiv_{obs(\mathcal{O})} \tilde{\mathcal{M}}$, *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 with respect to* $\mathcal{I} = \tilde{\mathcal{I}}$.

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, as the following example shows.

EXAMPLE 4.1.2 (Observational equivalence does not imply equivalence). *Consider the SCM* $\mathcal{M}$ *of Example 3.2.3 and let* $\tilde{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \mathbb{R}^2, \mathbb{R}^2, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ *be the SCM with the causal mechanism*

$$f_1(\boldsymbol{x}, \tilde{\boldsymbol{e}}) = \tilde{e}_1 \,, \quad f_2(\boldsymbol{x}, \tilde{\boldsymbol{e}}) = \gamma x_1 + \tilde{e}_2 \,,$$

where

$$\gamma = \frac{\beta \sigma_1^2 + \alpha \sigma_2^2}{\sigma_1^2 + \alpha^2 \sigma_2^2} \,,$$

and $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} = \mathbb{P}^{\tilde{\boldsymbol{E}}}$ *with* $\tilde{E}_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2)$, $\tilde{E}_2 \sim \mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2)$ *and* $\tilde{E}_1 \perp\!\!\!\perp \tilde{E}_2$, *where*

$$\begin{aligned} \tilde{\mu}_1 &= c[\mu_1 + \alpha \mu_2], & \tilde{\sigma}_1^2 &= c^2[\sigma_1^2 + \alpha^2 \sigma_2^2], \\ \tilde{\mu}_2 &= c[(\beta - \gamma)\mu_1 + (1 - \alpha\gamma)\mu_2], & \tilde{\sigma}_2^2 &= c^2[(\beta - \gamma)^2 \sigma_1^2 + (1 - \alpha\gamma)^2 \sigma_2^2] \end{aligned}$$

with $c = (1 - \alpha\beta)^{-1}$. *The augmented graph of* $\tilde{\mathcal{M}}$ *is depicted in Figure 6 (middle). The SCMs* $\mathcal{M}$ *and* $\tilde{\mathcal{M}}$ *are observationally equivalent, as one can check by explicit calculation. Similarly, one can define an SCM* $\hat{\mathcal{M}}$ *with augmented graph as depicted in Figure 6 (right) that is observationally equivalent to both* $\tilde{\mathcal{M}}$ *and* $\mathcal{M}$. *Because each of the SCMs have a different augmented graph, we conclude that none of the SCMs* $\mathcal{M}$, $\tilde{\mathcal{M}}$ *and* $\hat{\mathcal{M}}$ *are equivalent to each other.*

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. Although the SCMs of this example are all observationally equivalent, they are not interventionally equivalent, as we will see in the next section.

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

DEFINITION 4.2.1 (Interventional equivalence). *Consider two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are said to be* interventionally equivalent with respect to $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$*, denoted by $\mathcal{M} \equiv_{int(\mathcal{O})} \tilde{\mathcal{M}}$, 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 $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ and $\tilde{\mathcal{M}}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ are observationally equivalent with respect to $\mathcal{O}$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are called* interventionally equivalent *if they are interventionally equivalent with respect to $\mathcal{I} = \tilde{\mathcal{I}}$.*

Equivalent SCMs have the same solutions under every perfect intervention, and hence they are interventionally equivalent w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$.

It is clear that none of the three SCMs in Example 4.1.2 are interventionally equivalent to each other, although they are all observationally equivalent to each other. The following example illustrates that interventional equivalence w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply interventional equivalence w.r.t. a strict superset $\mathcal{O} \subsetneq \mathcal{Q} \subseteq \mathcal{I}$.

EXAMPLE 4.2.2 (Interventional equivalence is not preserved under taking a strict superset). *Consider the SCM $\mathcal{M}$ from Example 3.3.1 and consider the SCM $\tilde{\mathcal{M}} = \langle \mathbf{2}, \emptyset, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ with the following causal mechanism*

$$\tilde{f}_1(\boldsymbol{x}) = x_1(1 - \mathbf{1}_{\{1\}}(x_2)) - x_2 + 1\,, \quad \tilde{f}_2(\boldsymbol{x}) = x_2(1 - \mathbf{1}_{\{0\}}(x_1)) + 1\,.$$

Both SCMs are uniquely solvable and they are observationally equivalent. For every perfect intervention $\mathrm{do}(\{1\}, \xi_1)$ *with $\xi_1 \in \mathbb{R}$ their intervened models are observationally equivalent, and hence $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent w.r.t. $\{1\}$. Doing a perfect intervention* $\mathrm{do}(\{2\}, 1)$ *on both models also yields observationally equivalent SCMs. However, doing a perfect intervention* $\mathrm{do}(\{2\}, \xi_2)$ *for some $\xi_2 > 1$ leads for both models to a non-uniquely solvable SCM, for which $\mathcal{M}_{\mathrm{do}(\{2\},\xi_2)}$ is solvable and $\tilde{\mathcal{M}}_{\mathrm{do}(\{2\},\xi_2)}$ is non-solvable. Hence $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are not interventionally equivalent.*

Interventional equivalence w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ implies in particular observational equivalence w.r.t. $\mathcal{O}$, since the empty perfect intervention ($I = \emptyset$) is a special case of a perfect intervention. We saw already in Examples 4.1.2 and 4.2.2 that interventional equivalence is a strictly finer notion than observational equivalence. Even for this finer notion of interventional 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, as is shown in the following example.

EXAMPLE 4.2.3 (Interventionally equivalent SCMs with different graphs). *Consider the SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{2}, \{-1,1\}^2, \{-1,1\}^2, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ where the causal mechanism is given by*

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = e_1\,, \quad f_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 e_2$$

and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^{\boldsymbol{E}}$ with $E_1, E_2 \sim \mathcal{U}(\{-1,1\})$ uniformly distributed and $E_1 \perp\!\!\!\perp E_2$. In addition, consider the SCM $\tilde{\mathcal{M}}$ that is the same as $\mathcal{M}$ except for its causal mechanism, which is given by

$$\tilde{f}_1(\boldsymbol{x}, \boldsymbol{e}) = e_1\,, \quad \tilde{f}_2(\boldsymbol{x}, \boldsymbol{e}) = e_2\,.$$

Then $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent although $\mathcal{G}^a(\mathcal{M})$ is not equal to $\mathcal{G}^a(\tilde{\mathcal{M}})$, as can be seen in Figure 7.

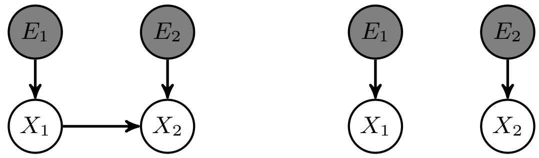

Fig 7: *The augmented graph of the SCM $\mathcal{M}$ (left) and of $\tilde{\mathcal{M}}$ (right) of Example 4.2.3.*

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 4.3.1 (Counterfactual equivalence). *Consider two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are said to be* counterfactually equivalent with respect to *$\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, denoted by $\mathcal{M} \equiv_{cf(\mathcal{O})} \tilde{\mathcal{M}}$, 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 finer than that of equivalence.

PROPOSITION 4.3.2. *Equivalent SCMs are counterfactually equivalent w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$.*

The counterfactual equivalence relation is related to the interventional equivalence relation in the following way.

PROPOSITION 4.3.3. *If two SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are counterfactually equivalent w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, then $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent w.r.t. $\mathcal{O}$.*

However, the converse is not true in general, as the next simple example shows.

EXAMPLE 4.3.4 (Interventional equivalence does not imply counterfactual equivalence). *Consider the same SCMs as in Example 4.2.3. We saw in that example that they are interventionally equivalent. However, they are not counterfactually equivalent, as $\mathcal{M}^{\mathrm{twin}}_{\mathrm{do}(\{1',1\},(1,-1))}$ is not observationally equivalent to $\tilde{\mathcal{M}}^{\mathrm{twin}}_{\mathrm{do}(\{1',1\},(1,-1))}$. To see this, consider the counterfactual query $p(X_{2'} = 1 | \mathrm{do}(X_{1'} = 1), \mathrm{do}(X_1 = -1), X_2 = 1)$. Both SCMs will give a different answer and hence $\mathcal{M}$ and $\tilde{\mathcal{M}}$ cannot be counterfactually equivalent.*

Even interventionally equivalent SCMs with the same causal mechanism (that differ only in their exogenous distribution) may not be counterfactually equivalent. For example, the SCMs $\mathcal{M}_\rho$ and $\mathcal{M}_{\rho'}$ with $\rho \neq \rho'$ in the following example (due to Dawid (2002)) are interventionally but not counterfactually equivalent.

EXAMPLE 4.3.5 (Counterfactual density unidentifiable from observational and interventional densities (Dawid, 2002)). *Let $\rho \in \mathbb{R}$ and*

$$\mathcal{M}_\rho = \langle \mathbf{2}, \mathbf{2}, \{0,1\} \times \mathbb{R}, \{0,1\} \times \mathbb{R}^2, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$$

be the SCM with causal mechanism given by

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = e_1 \,, \quad f_2(\boldsymbol{x}, \boldsymbol{e}) = e_{21}(1 - x_1) + e_{22} x_1$$

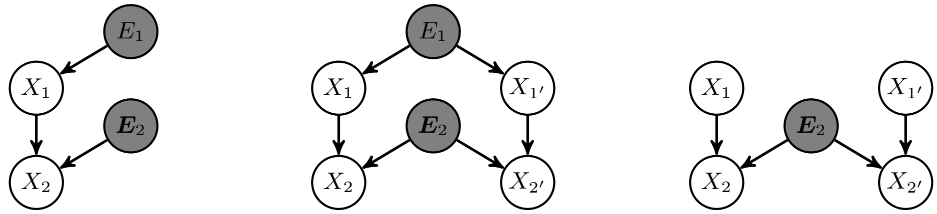

Fig 8: *The augmented graph of the SCM $\mathcal{M}_\rho$ (left), its twin SCM $\mathcal{M}_\rho^{\text{twin}}$ (center) and the intervened twin SCM $(\mathcal{M}_\rho^{\text{twin}})_{\text{do}(1,0),\text{do}(1',1)}$ (right) of Example 4.3.5.*

and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^{(E_1,\boldsymbol{E}_2)}$ *with* $E_1 \sim \text{Bernoulli}(1/2)$,

$$\boldsymbol{E}_2 := \begin{pmatrix} E_{21} \\ E_{22} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

normally distributed and $E_1 \perp\!\!\!\perp \boldsymbol{E}_2$. In an epidemiological setting, this SCM could be used to model whether a patient was treated or not (X_1) and the corresponding outcome for that patient (X_2).

Suppose in the actual world we did not assign treatment to a unit ($X_1 = 0$) and the outcome was $X_2 = c \in \mathbb{R}$. Consider the counterfactual query "What would the outcome have been, if we had assigned treatment to this unit?". We can answer this question by introducing a parallel counterfactual world that is modeled by the twin SCM $\mathcal{M}_\rho^{\text{twin}}$, as depicted in Figure 8. The counterfactual query then asks for $p(x_{2'} \,|\, \text{do}(X_{1'} = 1), \text{do}(X_1 = 0), X_2 = c)$. One can calculate that

$$p\left(\begin{pmatrix} x_{2'} \\ x_2 \end{pmatrix} \,|\, \text{do}(X_{1'} = 1), \text{do}(X_1 = 0)\right) = \mathcal{N}\left(\begin{pmatrix} x_{2'} \\ x_2 \end{pmatrix} \,\Big|\, \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

and hence $p(x_{2'} \,|\, \text{do}(X_{1'} = 1), \text{do}(X_1 = 0), X_2 = c) = \mathcal{N}(x_{2'} \,|\, -\rho c, 1 - \rho^2)$. Note that the answer to the counterfactual query depends on a quantity ρ that we cannot identify from the observational density $p(X_1, X_2)$ or the interventional densities $p(X_2 \,|\, \text{do}(X_1 = 0))$ and $p(X_2 \,|\, \text{do}(X_1 = 1))$, none of which depends on ρ. Therefore, even data from randomized controlled trials combined with observational data would not suffice to determine the value of this particular counterfactual query.

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.

Example 4.3.6 (Counterfactually equivalent SCMs with different graphs). *Consider the SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{2}, \{-1,1\}^2, \{-1,1\}^3, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with the causal mechanism $\boldsymbol{f}(\boldsymbol{x}, e_1, \boldsymbol{e}_2) := (e_1, e_{22})$ where $\boldsymbol{e}_2 = (e_{21}, e_{22}) \in \{-1,1\}^2$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^{(E_1,\boldsymbol{E}_2)}$ with $E_1, E_{21}, E_{22} \sim \mathcal{U}(\{-1,1\})$ independent. Let $\tilde{\mathcal{M}}$ be the same SCM as $\mathcal{M}$, but with the causal mechanism $\tilde{\boldsymbol{f}}(\boldsymbol{x}, e_1, \boldsymbol{e}_2) := (e_{21}, e_{22})$. Then $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are counterfactually equivalent (and, in particular, interventionally and observationally equivalent) although $1 \leftrightarrow 2 \in \mathcal{G}(\tilde{\mathcal{M}})$ but $1 \leftrightarrow 2 \notin \mathcal{G}(\mathcal{M})$.*

4.4. *Relations between equivalences.* The definitions of observational, interventional and counterfactual equivalence provide equivalence relations on the set of all SCMs. Note that for two SCMs to be observationally, interventionally or counterfactually equivalent w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, the endogenous spaces 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 set of the endogenous and the exogenous variables, the space of the endogenous and exogenous variables, the causal mechanism and the exogenous probability measure may all differ. The set of 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 Propositions 4.3.2 and 4.3.3):

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

This hierarchy formally establishes the "ladder of causation" (Shpitser and Pearl, 2008; Pearl and Mackenzie, 2018; Pearl, 2009) and allows us to compare SCMs at different levels of abstractions.

5. Marginalizations. In this section we show how, and under which condition, one can marginalize an SCM over a subset $\mathcal{L} \subsetneq \mathcal{I}$ 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.

In the literature the problem of marginalization of graphical models has been addressed for acyclic graph structures, e.g., ADMGs and mDAGs (see Verma, 1993; Richardson, 2003; Richardson and Spirtes, 2002; Evans, 2016, 2018, a.o.), and more recently in (Forré and Mooij, 2017) 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 under the *modularity* assumption on the underlying model (Forré and Mooij, 2017). We show that without this modularity assumption one can still define a marginalization for SCMs.

Intuitively, the idea is that we would like to treat the subsystem of the endogenous variables $\mathcal{L}$, described by the structural equations of the variables $\mathcal{L}$, as a "black box", and only describe how the rest of the system interacts with it. From Section 3.2 we know that if the SCM is uniquely solvable w.r.t. $\mathcal{L}$, then for each input value of the subsystem there exists an essentially unique assignment to an output value of the subsystem, which is described by a measurable solution function. By replacing the causal mechanism of this subsystem of the variables $\mathcal{L}$ with this measurable solution function, we completely remove the representation of the internals of the subsystem, and replace it with its essential input-output characteristics. We can thus abstract the subsystem away from the model by substituting the measurable solution function into the rest of the model, hereby effectively removing the functional dependence on the endogenous variables $\mathcal{L}$ from the rest of the system. An important property of this marginalization operation is that it preserves the observational, causal and the counterfactual semantics, meaning that the observational, interventional and counterfactual distributions induced by the SCM are on the margin identical to those induced by its marginalization.

Analogous to the marginalization of an SCM we define the marginalization of a directed mixed graph, which we call the latent projection, following standard terminology (Verma, 1993; Tian, 2002; Evans, 2016). We show that in general the marginalization of an SCM does not respect the latent projection of its associated (augmented) graph, i.e., that the (augmented) graph of the marginal SCM is not always a subgraph of the latent projection of the (augmented) graph of the original SCM. We show that under certain stronger ancestral unique solvability conditions the marginalization does respect the latent projection. The class of simple SCMs is closed under marginalization, and their marginalization always respects 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 5.1.1 (No SCM on the margin preserves the causal semantics). *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *with causal mechanism*

$$f_1(\boldsymbol{x}) = x_1 + x_2 + x_3\,, \quad f_2(\boldsymbol{x}) = x_2\,, \quad f_3(\boldsymbol{x}) = 0\,.$$

Then there exists no SCM $\tilde{\mathcal{M}}$ *on the 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 $\mathcal{L} \subsetneq \mathcal{I}$ one can never find an SCM $\tilde{\mathcal{M}}$ on the endogenous variables $\mathcal{I} \setminus \mathcal{L}$ that is interventionally equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.

We show next that the condition of unique solvability w.r.t. a certain subset is a sufficient condition for the existence of an SCM on the margin that preserves the causal semantics. Consider first the following example.

EXAMPLE 5.1.2 (SCM on the margin that preserves the causal semantics). *Consider the SCM* $\mathcal{M} = \langle \mathbf{5}, \mathbf{3}, \mathbb{R}^5, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$ *of Example 2.2.3. We saw that* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{L} = \{2,3,4\}$ *and that the mapping* $\boldsymbol{g}_{\mathcal{L}}$ *given in Example 3.1.2 is a measurable solution function. The system of structural equations for the variables* $\mathcal{L}$ *can be seen as a subsystem, that is, 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 every* $\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}$ *the input* $(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})})$ *gives these equations a unique output* $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}$*. This subsystem is depicted by the gray box in Figure 5. Substituting the components* $(g_{\mathcal{L}})_2, (g_{\mathcal{L}})_3$ *and* $(g_{\mathcal{L}})_4$ *into the causal mechanism components* f_1, f_5 *for the remaining endogenous variables* $\{1,5\}$ *gives a "marginal" causal mechanism*

$$\tilde{f}_1(\boldsymbol{x},\boldsymbol{e}) := x_1 - \alpha x_1^2 + e_1^2\,, \quad \tilde{f}_5(\boldsymbol{x},\boldsymbol{e}) := x_1 \cdot e_3 + e_1 \cdot e_3 + 2e_2 \cdot e_3\,.$$

These mappings define an SCM $\tilde{\mathcal{M}} := \langle \mathbf{2}, \mathbf{3}, \mathbb{R}^2, \mathbb{R}^3, \tilde{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^3} \rangle$ *on the margin* $\mathcal{I} \setminus \mathcal{L} = \{1,5\}$*. This constructed SCM* $\tilde{\mathcal{M}}$*, depicted in Figure 5, is interventionally equivalent w.r.t.* $\mathcal{L}$*, which can be checked manually or by applying Theorem 5.1.7.*

In general, for an SCM $\mathcal{M}$ and a given subset $\mathcal{L} \subsetneq \mathcal{I}$ of endogenous variables and its complement $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$, the structural equations $\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e})$ define a "subsystem". 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 $\boldsymbol{e}_{\mathrm{pa}(\mathcal{L})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})}$. We can effectively 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 5.1.7.

Definition 5.1.3 (Marginalization of an SCM). *Consider an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and a subset* $\mathcal{L} \subsetneq \mathcal{I}$*, and let* $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$*. Assume that* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{L}$*. 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* $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ *defined by*

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

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.* $\mathcal{L}$*.*

Note that for a specific $\mathcal{L} \subsetneq \mathcal{I}$ there may exist more than one marginalization w.r.t. $\mathcal{L}$, depending on the choice of the measurable solution function $\boldsymbol{g}_{\mathcal{L}}$. However, all marginalizations w.r.t. $\mathcal{L}$ map $\mathcal{M}$ to a representative of the same equivalence class $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ of SCMs. Moreover, marginalizing two equivalent SCMs w.r.t. $\mathcal{L}$ yield two equivalent marginal SCMs. Thus, the mapping $\mathrm{marg}(\mathcal{L})$ induces a well-defined mapping between the equivalence classes of SCMs.

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 3.2.2). Moreover, this definition extends that notion to certain cyclic SCMs, namely those that are uniquely solvable w.r.t. a certain subset. For linear SCMs this translates to a certain matrix invertibility condition which can be substituted into the causal mechanism to yield another linear marginal SCM (see Proposition B.5 in the Appendix B).

In general, marginalization is not always defined for all subsets. For instance, the SCM of Example 3.4.3 cannot be marginalized over the variable 2, but can be marginalized over the variables 1 and 2 together. The fact that we cannot marginalize over the single variable 2 in that example is due to the existence of a self-cycle at variable 2. It follows from Proposition 3.2.4 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 2.2.3. 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 5.1.4. *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.* $\mathcal{L}_1 \subsetneq \mathcal{I}$*. Let* $\mathcal{L}_2 \subsetneq \mathcal{I}$ *be 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* $\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1)(\mathcal{M}) \equiv \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{M})$*.*

Note that in the previous 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 the operations of intervention and marginalization commute.

Proposition 5.1.5. *Given an SCM* $\mathcal{M}$*, a subset* $\mathcal{L} \subsetneq \mathcal{I}$ *such that* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{L}$*, a subset* $I \subseteq \mathcal{I} \setminus \mathcal{L}$ *and a value* $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$*. Marginalization* $\mathrm{marg}(\mathcal{L})$ *commutes with perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$*, i.e.,* $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M}) \equiv (\mathrm{do}(I, \boldsymbol{\xi}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$*.*

Similarly, we show that the twin operation commutes with the marginalization operation.

PROPOSITION 5.1.6. *Let $\mathcal{M}$ be an SCM and $\mathcal{L} \subsetneq \mathcal{I}$ a subset such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$. Marginalization* $\operatorname{marg}(\mathcal{L})$ *commutes with the twin operation, i.e.,* $(\operatorname{marg}(\mathcal{L} \cup \mathcal{L}') \circ \operatorname{twin})(\mathcal{M}) \equiv (\operatorname{twin} \circ \operatorname{marg}(\mathcal{L}))(\mathcal{M})$*, where $\mathcal{L}'$ is the copy of $\mathcal{L}$ in $\mathcal{I}'$.*

With Propositions 5.1.5 and 5.1.6 at hand we can prove the main result of this section.

THEOREM 5.1.7 (Marginalization of an SCM preserves the observational, causal and counterfactual semantics). *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L} \subsetneq \mathcal{I}$ such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$. Then $\mathcal{M}$ and* $\operatorname{marg}(\mathcal{L})(\mathcal{M})$ *are observationally, interventionally and counterfactually equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.*

As we saw in Example 4.3.4 it is generally not true that interventional equivalence implies counterfactual equivalence. 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.

We would like to stress that for an SCM $\mathcal{M}$, unique solvability w.r.t. a certain subset $\mathcal{L} \subsetneq \mathcal{I}$, 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. $\mathcal{I} \setminus \mathcal{L}$. This is illustrated by the following example.

EXAMPLE 5.1.8 (Marginalization condition of an SCM is not a necessary condition). *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*

$$f_1(\boldsymbol{x}, e) = e\,, \quad f_2(\boldsymbol{x}, e) = x_1\,, \quad f_3(\boldsymbol{x}, e) = x_2\,, \quad f_4(\boldsymbol{x}, e) = x_4$$

and $\mathbb{P}_{\mathbb{R}}$ is the standard-normal measure on $\mathbb{R}$. This SCM is solvable w.r.t. $\mathcal{L} = \{2, 4\}$, but not uniquely solvable w.r.t. $\mathcal{L}$, and hence we cannot apply Definition 5.1.3 to $\mathcal{L}$. However, the SCM $\tilde{\mathcal{M}}$ on the endogenous variables $\{1, 3\}$ with the causal mechanism $\tilde{\boldsymbol{f}}$ given by

$$\tilde{f}_1(\boldsymbol{x}, e) = e\,, \quad \tilde{f}_3(\boldsymbol{x}, e) = x_1$$

is counterfactually equivalent to $\mathcal{M}$ w.r.t. $\{1, 3\}$, which can be checked easily.

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". The name "latent projection" is inspired from a similar construction on directed mixed graphs in (Verma, 1993). In (Verma, 1993), the authors concentrate on a mapping between directed mixed graphs and show that it preserves conditional independence properties (see also Tian, 2002). In this section, we provide a sufficient condition for the marginalization of an SCM to respect the latent projection, i.e., that the augmented graph of the marginal SCM is a subgraph of the latent projection of the augmented graph of the original SCM.

DEFINITION 5.2.1 (Marginalization of a directed mixed graph). *Given a directed mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *and a subset* $\mathcal{L} \subseteq \mathcal{V}$*, we define the* marginalization of $\mathcal{G}$ w.r.t. $\mathcal{L}$ *or the* latent projection of $\mathcal{G}$ onto $\mathcal{V} \setminus \mathcal{L}$ *as the 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. $i \to j \in \tilde{\mathcal{E}}$ *iff there exists a directed path* $i \to \ell_1 \to \cdots \to \ell_n \to j$ *in* $\mathcal{G}$ *with* $n \geq 0$ *and* $\ell_1, \ldots, \ell_n \in \mathcal{L}$,
3. $i \leftrightarrow j \in \tilde{\mathcal{B}}$ *iff*
   (a) *there exist* $n, m \geq 0$, $\ell_1, \ldots, \ell_n \in \mathcal{L}$, $\tilde{\ell}_1, \ldots, \tilde{\ell}_m \in \mathcal{L}$ *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* $n, m \geq 1$, $\ell_1, \ldots, \ell_n \in \mathcal{L}$, $\tilde{\ell}_1, \ldots, \tilde{\ell}_m \in \mathcal{L}$ *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, note that if for a subgraph $\mathcal{H} \subseteq \mathcal{G}$, $\mathrm{marg}(\mathcal{L})(\mathcal{H}) \subseteq \mathrm{marg}(\mathcal{L})(\mathcal{G})$ for any subset of nodes $\mathcal{L}$. It does not matter in which order we project out the nodes or if we do several at once.

PROPOSITION 5.2.2. *Given a directed mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *and two disjoint subsets* $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{V}$*, then* $(\mathrm{marg}(\mathcal{L}_1) \circ \mathrm{marg}(\mathcal{L}_2))(\mathcal{G}) = (\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1))(\mathcal{G}) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{G})$.

Similarly 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 5.2.3. *Given a directed mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *and subsets* $\mathcal{L}, I \subseteq \mathcal{V}$*. If* $\mathcal{L}$ *and* I *are disjoint, then* $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I))(\mathcal{G}) = (\mathrm{do}(I) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})$.

PROPOSITION 5.2.4. *Given a directed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *and subsets* $\mathcal{L}, I \subseteq \mathcal{V}$ *such that* $\mathrm{pa}_{\mathcal{G}}(\mathcal{V} \setminus I) = \emptyset$*. If* $\mathcal{L} \subseteq I$*, then* $(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin}(I))(\mathcal{G}) = (\mathrm{twin}(I \setminus \mathcal{L}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})$*, where* $\mathcal{L}'$ *is the copy of* $\mathcal{L}$ *in* I'.

In Example 5.1.2 we already saw an example of a marginalization that respects the latent projection. However, not all marginalizations respect the latent projection, as is illustrated in the following example.

EXAMPLE 5.2.5 (Marginalization does not respect the latent projection). *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*

$$\begin{aligned} f_1(\boldsymbol{x}, e) &= x_1(1 - \mathbf{1}_{\{0\}}(x_2 - x_3)) + 1\,, & f_2(\boldsymbol{x}, e) &= x_2\,, \\ f_3(\boldsymbol{x}, e) &= e\,, & f_4(\boldsymbol{x}, e) &= x_2 \end{aligned}$$

and $\mathbb{P}_{\mathbb{R}}$ *the standard-normal measure on* $\mathbb{R}$*. Although* $\mathcal{M}$ *and its marginalization* $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ *with* $\mathcal{L} = \{1, 2\}$ *are interventionally equivalent w.r.t.* $\mathcal{I} \setminus \mathcal{L} = \{3, 4\}$*, the augmented graph* $\mathcal{G}^a(\mathcal{M}_{\mathrm{marg}(\mathcal{L})})$ *is not a subgraph of the latent projection of* $\mathcal{G}^a(\mathcal{M})$ *onto* $\mathcal{I} \setminus \mathcal{L}$*, as can be verified from the augmented graphs depicted in Figure 9.*

Under the stronger ancestral unique solvability condition one can prove that the marginalization of an SCM respects the latent projection.

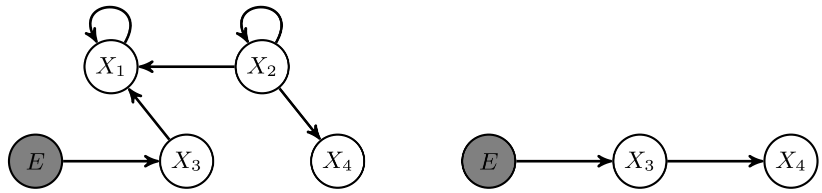

Fig 9: *The augmented graph of the SCM $\mathcal{M}$ (left) and of $\mathcal{M}_{\mathrm{marg}(\{1,2\})}$ (right) as in Example 5.2.5.*

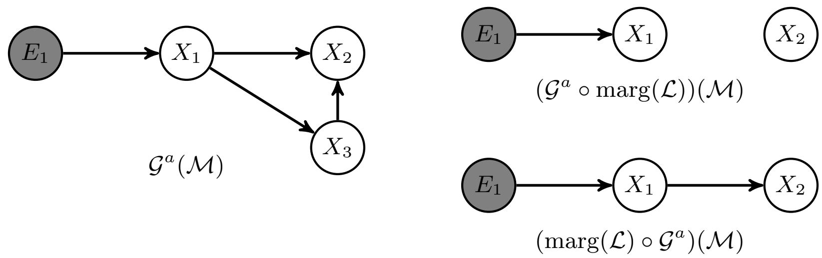

Fig 10: *Example of a marginalization of $\mathcal{M}$ w.r.t. $\mathcal{L} = \{3\}$ for which the augmented graph is a strict subgraph of the latent projection of $\mathcal{G}^a(\mathcal{M})$ onto $\mathcal{I} \setminus \mathcal{L}$, as described in Example 5.2.7.*

PROPOSITION 5.2.6. *Let $\mathcal{M}$ be an SCM that is ancestrally uniquely solvable w.r.t. $\mathcal{L} \subsetneq \mathcal{I}$. Then $\big(\mathcal{G}^a \circ \mathrm{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\mathrm{marg}(\mathcal{L}) \circ \mathcal{G}^a\big)(\mathcal{M})$ and $\big(\mathcal{G} \circ \mathrm{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\mathrm{marg}(\mathcal{L}) \circ \mathcal{G}\big)(\mathcal{M})$.*

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 B.4 in the Appendix B), and hence, a marginalization of a linear SCM always respects the latent projection.

The following example illustrates why the (augmented) graph of a marginalized SCM can be a strict subgraph of the corresponding latent projection.

EXAMPLE 5.2.7 (Graph of the marginal SCM is a strict subgraph of the latent projection). *Consider the SCM given by $\mathcal{M} = \langle \mathbf{3}, \mathbf{1}, \mathbb{R}^3, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$, where*

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = e_1 , \quad f_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 - x_3 , \quad f_3(\boldsymbol{x}, \boldsymbol{e}) = x_1$$

and take for $\mathbb{P}_{\mathbb{R}}$ the standard-normal measure on $\mathbb{R}$. Marginalizing over $\{3\}$ gives us the marginal causal mechanism

$$\tilde{f}_1(\boldsymbol{x}, \boldsymbol{e}) = e_1 , \quad \tilde{f}_2(\boldsymbol{x}, \boldsymbol{e}) = 0 .$$

Here, we see that the causal mechanism $\tilde{f}_2$ does not not depend on x_1, contrary to what one would expect from the latent projection (see Figure 10).

We finish this section with the following result.

PROPOSITION 5.2.8. *Simplicity is preserved under marginalization.*

That is, on simple SCMs one can perform any number of marginalizations in any order. Moreover, all these marginalizations respect the latent projection (see Proposition 5.2.6) and each resulting marginal SCM is simple and thus in particular has no self-cycles.

From Proposition 2.4.4 and Propositions 3.2.2 and 5.2.6 we recover the known result that the class of acyclic SCMs is closed under both perfect intervention and marginalization (Evans, 2016). Similarly, from Proposition 3.5.2 and 5.2.8 it follows that the class of simple SCMs is also closed under both perfect intervention and marginalization. This makes both classes convenient to work with.

6. Markov properties. In this section we give a short overview of Markov properties for SCMs with cycles. We will make use of the Markov properties that were recently developed by Forré and Mooij (2017) 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. We also provide a shorter and more intuitive derivation so that this section can act as an entry point for the reader into the far more extensive and detailed discussion of Markov properties that Forré and Mooij (2017) give.

Markov properties associate a set of conditional independence relations to a graph. The *directed global Markov property* for directed acyclic graphs, also known as the "d-separation criterion" (Pearl, 1985), is one of the most widely used. It directly extends to a similar property for acyclic directed mixed graphs (ADMGs) (Richardson, 2003). It does not hold in general for cyclic SCMs, however, as was already observed earlier (Spirtes, 1994, 1995). Under some conditions (roughly speaking, linearity or discrete variables) the directed global Markov property can be shown to hold also in the presence of cycles (Forré and Mooij, 2017).

Inspired by work of Spirtes (1994), Forré and Mooij (2017) also recognized that in the general cyclic case a different extension of d-separation, termed σ-separation, is needed, leading to the *general directed global Markov property*. One key result in (Forré and Mooij, 2017) implies that under the assumption of unique solvability w.r.t. each strongly connected component of its graph, the observational distribution of an SCM satisfies the general directed global Markov property w.r.t. its graph. The solvability assumptions are in general not preserved under interventions. Under the stronger assumption of simplicity, however, they are, and one obtains the corollary that also all interventional and counterfactual distributions of a simple SCM satisfy the general directed global Markov property w.r.t. to their corresponding graphs.

For a more extensive study of different Markov properties that can be associated to SCMs we refer the reader to (Forré and Mooij, 2017). In Appendix A one can find all the relevant graphical notation and terminology for the remainder of this section.

6.1. *The directed global Markov property.* A convenient way to read off conditional independencies from the graph of an acyclic SCM is by using the graphical criterion called d-separation (Pearl, 2009). The directed global Markov property associates a conditional independence relation in the observational distribution of the SCM to each d-separation entailed by the graph. Here, we use a formulation of d-separation that generalizes d-separation for DAGs (Pearl, 1985) and m-separation for ADMGs (Richardson, 2003) and mDAGs (Evans, 2016), as discussed in more detail in (Forré and Mooij, 2017). We refer the reader to Appendix A.2 for the definitions of d-separation (Definition A.2.2) and the directed global Markov property (Definition A.2.4).

From the results in (Forré and Mooij, 2017) it directly follows that for the observational distribution of an SCM, the directed global Markov property w.r.t. the graph of the SCM holds under one of the following assumptions.

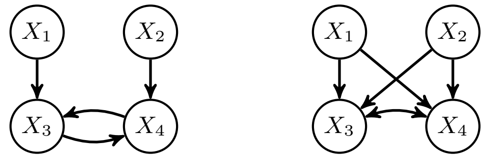

Fig 11: *The graphs of the observationally equivalent SCMs $\mathcal{M}$ (left) and $\tilde{\mathcal{M}}$ (right) of Example 6.1.2 and 6.2.1.*

THEOREM 6.1.1 (d-separation criterion for SCMs (Forré and Mooij, 2017)). *Consider a uniquely solvable SCM $\mathcal{M}$ that satisfies at least one of the following three conditions:*

1. *$\mathcal{M}$ is acyclic;*
2. *all endogenous spaces $\mathcal{X}_i$ are discrete and $\mathcal{M}$ is ancestrally uniquely solvable;*
3. *$\mathcal{M}$ is linear (see Definition B.1 in the Appendix B), each of its causal mechanisms $\{f_i\}_{i\in\mathcal{I}}$ has a non-trivial dependence on at least one exogenous variable, and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{\mathcal{J}}$.*

Then, its observational distribution $\mathbb{P}^{\boldsymbol{X}}$ exists, is unique and obeys the directed global Markov property relative to $\mathcal{G}(\mathcal{M})$ (see Definition A.2.4).

The acyclic case is well-known and was first shown in the context of linear-Gaussian structural equation models (Spirtes et al., 1998; Koster, 1999). The discrete case fixes the erroneous theorem by Pearl and Dechter (1996), for which a counterexample was found by Neal (2000), 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 (Spirtes, 1994, 1995; Koster, 1996) to a linear (possibly non-Gaussian) setting with bidirected edges in the graph.

The following counterexample of an SCM for which the directed global Markov property does not hold was already given in (Spirtes, 1994, 1995).

EXAMPLE 6.1.2 (Directed global Markov property does not hold for cyclic SCM). *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*

$$f_1(\boldsymbol{x},\boldsymbol{e}) = e_1\,,\ f_2(\boldsymbol{x},\boldsymbol{e}) = e_2\,,\ f_3(\boldsymbol{x},\boldsymbol{e}) = x_1x_4 + e_3\,,\ f_4(\boldsymbol{x},\boldsymbol{e}) = x_2x_3 + e_4$$

and $\mathbb{P}_{\mathbb{R}^4}$ is the standard-normal distribution on $\mathbb{R}^4$. The graph of $\mathcal{M}$ is depicted in Figure 11 on the left. The model is uniquely solvable (it is even simple). One can check that for every solution $\boldsymbol{X}$ of $\mathcal{M}$, 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 $\mathcal{G}(\mathcal{M})$. Hence the global directed Markov property does not hold here.

6.2. *The general directed global Markov property.* In (Forré and Mooij, 2017), the notion of "acyclification" is introduced. This is an alternative to the "collapsed graph" for directed graphs of (Spirtes, 1994) that is applicable to 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, the causal mechanisms of these strongly connected components can be replaced by their measurable solution functions, which results in an acyclic SCM. This acyclification preserves the solutions, and d-separation in the acyclification can directly be translated into "σ-separation" on the original graph. This then leads to the general directed global Markov property. We will discuss this now in more detail.

EXAMPLE 6.2.1 (Construction of an observationally equivalent acyclic SCM). *Consider the SCM $\mathcal{M}$ of Example 6.1.2. The mappings*

$$g_3(x_1, x_2, \boldsymbol{e}) := \tfrac{x_1 e_4 + e_3}{1 - x_1 x_2}, \quad g_4(x_1, x_2, \boldsymbol{e}) := \tfrac{x_2 e_3 + e_4}{1 - x_1 x_2}$$

are measurable solution functions for $\mathcal{M}$ w.r.t. $\{3,4\}$. One can easily check that $\mathcal{M}$ is uniquely solvable w.r.t. $\{3,4\}$. Consider now the SCM $\tilde{\mathcal{M}}$ that is the same as $\mathcal{M}$ except that its causal mechanism $\tilde{\boldsymbol{f}}$ is defined by $\tilde{\boldsymbol{f}} := (f_1, f_2, g_3, g_4)$. By construction, $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are observationally equivalent. Because $\tilde{\mathcal{M}}$ is acyclic (see Figure 11 on the right) 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 $\mathcal{M}$).

One of the key insights in (Forré and Mooij, 2017) is that this example can easily be generalized as follows.

DEFINITION 6.2.2 (Acyclification of an SCM). *For an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$ we call an SCM $\mathcal{M}^{\mathrm{acy}}$ an* acyclification of $\mathcal{M}$ *if $\mathcal{M}^{\mathrm{acy}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \hat{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with "acyclified" causal mechanism $\hat{\boldsymbol{f}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ given by*

$$\hat{f}_i(\boldsymbol{x}, \boldsymbol{e}) = g_i(\boldsymbol{x}_{\mathrm{pa}(\mathrm{sc}(i)) \setminus \mathrm{sc}(i)}, \boldsymbol{e}_{\mathrm{pa}(\mathrm{sc}(i))}), \quad i \in \mathcal{I},$$

where pa *and* sc *denote the parents and strongly connected components according to $\mathcal{G}^a(\mathcal{M})$, and where g_i is the i-component of a measurable solution function $\boldsymbol{g}_{\mathrm{sc}(i)} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathrm{sc}(i)) \setminus \mathrm{sc}(i)} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathrm{sc}(i))} \to \boldsymbol{\mathcal{X}}_{\mathrm{sc}(i)}$ of $\mathcal{M}$ w.r.t. $\mathrm{sc}(i)$. We denote by $\mathrm{acy}(\mathcal{M})$ the equivalence class of the acyclifications of $\mathcal{M}$.*

Note that $\mathrm{acy}(\mathcal{M})$ is well-defined: all acyclifications of an SCM $\mathcal{M}$ map $\mathcal{M}$ to a representative of the same equivalence class of SCMs.

PROPOSITION 6.2.3. *Consider an SCM $\mathcal{M}$ that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. Then an acyclification $\mathcal{M}^{\mathrm{acy}}$ of $\mathcal{M}$ is acyclic and observationally equivalent to $\mathcal{M}$.*

We can also define a graphical acyclification for directed mixed graphs, which is a special case of the operation defined in (Forré and Mooij, 2017) for HEDGes.

DEFINITION 6.2.4 (Acyclification of a directed mixed graph). *For a directed mixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$, the* acyclification of $\mathcal{G}$ *is defined by the directed mixed graph $\mathcal{G}^{\mathrm{acy}} = (\mathcal{V}, \hat{\mathcal{E}}, \hat{\mathcal{B}})$ with directed edges $j \to i \in \hat{\mathcal{E}}$ if and only if $j \in \mathrm{pa}_{\mathcal{G}}(\mathrm{sc}_{\mathcal{G}}(i)) \setminus \mathrm{sc}_{\mathcal{G}}(i)$ and bidirected edges $i \leftrightarrow j \in \hat{\mathcal{B}}$ if and only if there exist $i' \in \mathrm{sc}_{\mathcal{G}}(i)$ and $j' \in \mathrm{sc}_{\mathcal{G}}(j)$ with $i' = j'$ or $i' \leftrightarrow j' \in \mathcal{B}$.*

The following compatibility result is immediate from the definitions.

PROPOSITION 6.2.5. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. Then $\mathcal{G}^a(\mathrm{acy}(\mathcal{M})) \subseteq \mathrm{acy}(\mathcal{G}^a(\mathcal{M}))$ and $\mathcal{G}(\mathrm{acy}(\mathcal{M})) \subseteq \mathrm{acy}(\mathcal{G}(\mathcal{M}))$.*

The following example illustrates that the graph of an acyclified SCM can be a strict subgraph of the acyclification of the graph of the SCM.

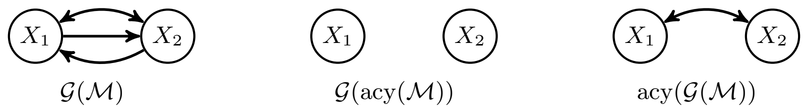

Fig 12: *The graphs of the original SCM $\mathcal{M}$ (left), of the acyclified SCM (centre), and of the acyclification of the graph of $\mathcal{M}$ (right) corresponding to Example 6.2.6.*

Example 6.2.6 (Graph of the acyclification of the SCM is a strict subgraph of the acyclification of its graph). *Consider the SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ with the causal mechanism defined by*

$$f_1(\boldsymbol{x}, e) = x_2 - e\,, \quad f_2(\boldsymbol{x}, e) = \tfrac{1}{2}x_1 + e$$

and $\mathbb{P}_{\mathbb{R}}$ the standard normal measure on $\mathbb{R}$. The SCM $\mathcal{M}$ is uniquely solvable w.r.t. the (only) strongly connected component $\{1,2\}$. An acyclification of $\mathcal{M}$ is the SCM $\mathcal{M}^{\mathrm{acy}}$ with acyclified causal mechanism $\hat{\boldsymbol{f}}$ defined by

$$\hat{f}_1(\boldsymbol{x}, e) = 0\,, \quad \hat{f}_2(\boldsymbol{x}, e) = e\,.$$

The graph $\mathcal{G}(\mathrm{acy}(\mathcal{M}))$ is a strict subgraph of $\mathrm{acy}(\mathcal{G}(\mathcal{M}))$ as can be seen in Figure 12.

Translating the notion of d-separation from the acyclified graph back to the original graph led to the notion of σ-separation.

Definition 6.2.7 (σ-separation (Forré and Mooij, 2017)). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and let $C \subseteq \mathcal{V}$ be a subset of nodes. A walk*[17] *(path) $\pi = (i_0, \epsilon_1, i_1, \dots, i_n)$ in $\mathcal{G}$ is said to be* C-σ-blocked or σ-blocked by C *if*

1. *its first node $i_0 \in C$ or its last node $i_n \in C$, or*
2. *it contains a collider $i_k \notin \mathrm{an}_{\mathcal{G}}(C)$, or*
3. *it contains a non-endpoint non-collider $i_k \in C$ that points towards a neighboring node on π that lies in a different strongly connected component of $\mathcal{G}$, i.e., such that $i_{k-1} \leftarrow i_k$ in π and $i_{k-1} \notin \mathrm{sc}_{\mathcal{G}}(i_k)$, or $i_k \to i_{k+1}$ in π and $i_{k+1} \notin \mathrm{sc}_{\mathcal{G}}(i_k)$.*

The walk (path) π is said to be C-σ-open *if it is not σ-blocked by C. For two subsets of nodes $A, B \subseteq \mathcal{V}$, we say that* A is σ-separated from B given C in $\mathcal{G}$ *if all paths between a node in A and a node in B are σ-blocked by C, and write*

$$A \underset{\mathcal{G}}{\overset{\sigma}{\perp}} B \,|\, C\,.$$

The only difference between σ-separation and d-separation is that d-separation does not have the extra condition on the non-collider that it has to point to a node in a different strongly connected component. It is therefore obvious that σ-separation reduces to d-separation for acyclic graphs, since $\mathrm{sc}_{\mathcal{G}}(i) = \{i\}$ for each $i \in \mathcal{V}$ in that case.

Although for proofs it is often easier to make use of walks, it suffices to formulate σ-separation in term of paths rather than walks because of the following result, which is analogous to a similar result for d-separation (see Lemma A.2.3 in the Appendix A.2).

[17] For the definitions of path, walk, collider and non-collider we refer the reader to Appendix A.2.

LEMMA 6.2.8. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph, $C \subseteq \mathcal{V}$ and $i, j \in \mathcal{V}$. There exists a C-σ-open walk between i and j in $\mathcal{G}$ if and only if there exists a C-σ-open path between i and j in $\mathcal{G}$.*

It is clear from the definitions that σ-separation implies d-separation. The other way around does not hold in general, as can be seen in the following example.

EXAMPLE 6.2.9 (d-separation does not imply σ-separation). *Consider the directed graph $\mathcal{G}$ as depicted in Figure 11 (left). Here X_1 is d-separated from X_2 given $\{X_3, X_4\}$, but X_1 is not σ-separated from X_2 given $\{X_3, X_4\}$.*

The following result relates σ-separation to d-separation.

PROPOSITION 6.2.10. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph. Then for $A, B, C \subseteq \mathcal{V}$,*

$$A \underset{\mathcal{G}}{\overset{\sigma}{\perp}} B \,|\, C \iff A \underset{\mathrm{acy}(\mathcal{G})}{\overset{d}{\perp}} B \,|\, C\,.$$

By replacing d-separation by σ-separation, one obtains the formulation of what Forré and Mooij (2017) termed the "generalized" directed global Markov property.

DEFINITION 6.2.11 (General directed global Markov property (Forré and Mooij, 2017)). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathbb{P}_{\mathcal{V}}$ a probability distribution on $\boldsymbol{\mathcal{X}}_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$. The probability distribution $\mathbb{P}_{\mathcal{V}}$ obeys the* general directed global Markov property *relative to $\mathcal{G}$ if for all subsets $A, B, C \subseteq \mathcal{V}$ we have*

$$A \underset{\mathcal{G}}{\overset{\sigma}{\perp}} B \,|\, C \quad \Longrightarrow \quad \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{V}}}{\perp\!\!\!\perp} \boldsymbol{X}_B \,|\, \boldsymbol{X}_C\,,$$

i.e., $(X_i)_{i \in A}$ and $(X_i)_{i \in B}$ are conditionally independent given $(X_i)_{i \in C}$ under $\mathbb{P}_{\mathcal{V}}$, where we take the canonical projections $X_i : \boldsymbol{\mathcal{X}}_{\mathcal{V}} \to \mathcal{X}_i$ as random variables.

The fact that σ-separation implies d-separation means that the directed global Markov property implies the general directed global Markov property. In other words, the general directed global Markov property is weaker than the directed global Markov property. It is actually strictly weaker, as we saw in Example 6.2.9.

The following fundamental result follows directly from the theory in (Forré and Mooij, 2017).

THEOREM 6.2.12 (σ-separation criterion for SCMs). *Consider an SCM $\mathcal{M}$ that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. Then its observational distribution $\mathbb{P}^{\boldsymbol{X}}$ exists, is unique and it obeys the general directed global Markov property relative to $\mathcal{G}(\mathcal{M})$.*[18]

[18] Since (Forré and Mooij, 2017) 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 6.2.12 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 $\mathcal{M} = \langle \mathbf{2}, \emptyset, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ 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 $\{2\}$. Since the graph contains no edges, one might naïvely expect that for every solution $\boldsymbol{X}$ we have $X_1 \perp\!\!\!\perp X_2$. However, that is clearly not true: the solution with $X_1 = X_2$, where X_2 has a non-degenerate distribution, shows a dependence between X_1 and X_2. In general, all strongly connected components that admit multiple solutions may be dependent on any other variable(s) in the model.

Intuitively, for $A, B, C \subseteq \mathcal{I}$, if A is σ-separated from B given C in $\mathcal{G}(\mathcal{M})$, then A is d-separated from B by C in $\mathrm{acy}(\mathcal{G}(\mathcal{M}))$ and hence in $\mathcal{G}(\mathrm{acy}(\mathcal{M}))$, and since $\mathrm{acy}(\mathcal{M})$ is acyclic and observationally equivalent to $\mathcal{M}$, it follows from the directed global Markov property applied to $\mathrm{acy}(\mathcal{M})$ that $\boldsymbol{X}_A \perp\!\!\!\perp_{\mathbb{P}^{\boldsymbol{X}}} \boldsymbol{X}_B \mid \boldsymbol{X}_C$ for every solution $\boldsymbol{X}$ of $\mathcal{M}$. Note that the ancestral unique solvability condition for the discrete case is strictly weaker than the condition of unique solvability w.r.t. each strongly connected component in Theorem 6.2.12. For the linear case, the condition of unique solvability is equivalent to the condition of unique solvability w.r.t. each strongly connected component (see Proposition B.4 in the Appendix B).

The results in Theorems 6.1.1 and 6.2.12 are not preserved under perfect intervention, because intervening on a strongly connected component could split it into several strongly connected components with different solvability properties. As the class of simple SCMs is preserved under perfect intervention and the twin operation (Proposition 3.5.2), we obtain the following corollary.

COROLLARY 6.2.13 (Directed global Markov properties for a simple SCM). *Consider a simple SCM $\mathcal{M}$. 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 obey 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), (2), (3) of Theorem 6.1.1, 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.

7. Causal interpretation of the graph of SCMs. In Examples 4.2.3 and 4.3.6 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 all the conditions under which the presence or absence of a directed path or bidirected edge in the graph can be identified from the observational and interventional distributions. Instead, we give some sufficient conditions under which one can detect a directed path and bidirected edge in the graph.

In general, cyclic SCMs may have multiple, one, or zero induced observational distributions, and this may change after intervention. Since it is not clear what an experimenter would measure in case of multiple induced distributions (how does nature choose from those?) or zero induced distributions (e.g., in case of an asymptotically oscillating system (Rubenstein et al., 2018)), we restrict ourselves here 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 5.2.5. Here, one sees a difference in the marginal distribution $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{3\},\xi_3)}}$ on $\mathcal{X}_4$ for different values of ξ_3, although variable 3 is not an ancestor of variable 4 and each marginal distribution $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{3\},\xi_3)}}$ on $\mathcal{X}_4$ is uniquely defined. This counterintuitive behavior that an intervention on a non-ancestor of a variable can change the distribution of that variable was already observed by Neal (2000). However, under a specific unique solvability condition, we obtain a direct causal interpretation for the absence of a direct edge or directed path in the graph of an SCM.

PROPOSITION 7.1.1 (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, i.e., there exists a measurable set* $\mathcal{B}_j \subseteq \mathcal{X}_j$ *such that*

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

then there exists a directed edge $i \to j$ *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 $\mathcal{G}(\mathcal{M})$. We carefully avoided making statements when an induced marginal distribution is not defined or not unique.[19]

We would like to stress that the condition in Proposition 7.1.1 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. For example, no interventional distribution satisfies the condition of Proposition 7.1.1 for the SCM $\mathcal{M}$ in Example 4.2.3, although there is a directed edge $1 \to 2$ in the graph $\mathcal{G}(\mathcal{M})$.

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 (Pearl, 2009). 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 $X_i = x_i$.

EXAMPLE 7.2.1 (Detecting a bidirected edge in the graph of an SCM). *Consider the acyclic SCM* $\mathcal{M}$ *of Example 4.2.3 and the SCM* $\hat{\mathcal{M}}$ *that is the same as* $\mathcal{M}$ *except for its causal mechanism, which is given by*

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = e_1\,, \quad f_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 e_1\,.$$

For the SCM $\hat{\mathcal{M}}$ *we observe that the marginal interventional probability* $\mathbb{P}_{\hat{\mathcal{M}}_{\mathrm{do}(\{1\},\xi_1)}}(X_2 = -1)$ *is not equal to the conditional probability* $\mathbb{P}_{\hat{\mathcal{M}}}(X_2 = -1 \mid X_1 = \xi_1)$ *for both* $\xi_1 = -1$ *and* $\xi_1 = 1$*. This observation suffices to identify the presence of the bidirected edge* $1 \leftrightarrow 2$ *in the graph* $\mathcal{G}(\hat{\mathcal{M}})$*. For the SCM* $\mathcal{M}$*, whose graph does not contain the bidirected edge* $1 \leftrightarrow 2$*, the marginal interventional distribution and conditional distribution coincide.*

The following proposition provides a generalization of this well-known result to SCMs that may include cycles.

[19] E.g. it would be tempting to propose the condition that there exists a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ and there exist values $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ such that $\left(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}\right)_{\mathrm{do}(i,\xi_i)}$ and $\left(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}\right)_{\mathrm{do}(i,\tilde{\xi}_i)}$ are not observationally equivalent with respect to $\{j\}$. However, it is not clear what this would mean in practice if the marginal observational distributions do not exist or are not unique. The assumptions of existence and uniqueness enable one to test in practice whether the two interventional distributions coincide by using two finite samples from both distributions.

PROPOSITION 7.2.2 (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*

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

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 4.3.6. There, we saw that for the SCM $\tilde{\mathcal{M}}$, there exists a bidirected edge $1 \leftrightarrow 2 \in \mathcal{G}(\tilde{\mathcal{M}})$ 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 also can 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). Unfortunately, it i and j are part of a cycle, the above sufficient condition cannot be applied, and currently no simple sufficient conditions for detecting the presence of a bidirected edge in that case are known.

7.3. *Simple SCMs.* Simple SCMs always satisfy the unique solvability conditions of Proposition 7.1.1 and 7.2.2, and moreover, the (marginal) observational and interventional distributions always exist and are uniquely defined. This allows us to define causal relationships for simple SCMs in terms of its graph.

DEFINITION 7.3.1 (Causal relationships of a simple SCM). *Let $\mathcal{M}$ be a simple SCM.*

1. *If there exists a directed edge $i \to j \in \mathcal{G}(\mathcal{M})$, i.e., $i \in \mathrm{pa}(j)$, 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 $\mathcal{G}(\mathcal{M})$, i.e., $i \in \mathrm{an}(j)$, 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 confounded according to *$\mathcal{M}$.*

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

COROLLARY 7.3.2 (Sufficient conditions for the presence of causal 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 $\mathcal{M}$, i.e., $i \to j \in \mathcal{G}(\mathcal{M})$;

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 $\mathcal{M}$, i.e., $i \to \cdots \to j$ 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*

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

then i and j are confounded according to $\mathcal{M}$, i.e., $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$.

We would like to stress that even for simple SCMs it is not possible to identify all the causal relationships in the graph from the observational, interventional, or even the counterfactual distributions. We saw already in Example 4.2.3 and 4.3.6 that this is impossible for acyclic SCMs without further assumptions.

8. 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. We saw that cyclic SCMs behave quite 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. We would like to emphasize that most of these results hold only under certain sufficient conditions that are not necessary (e.g., for the marginalization operation this was shown in Example 5.1.8). 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 (see Corollary 3.5.3); the class of simple SCMs is closed under both perfect intervention and marginalization (see Proposition 3.5.2 and 5.2.8); the marginalization respects the latent projection (see Proposition 5.2.6); the induced distributions obey the general directed global Markov property and obey the directed global Markov property in the acyclic, discrete and linear case (see Corollary 6.2.13). This class is free from SCMs that have self-cycles (see Proposition 3.2.4) and their graphs have a direct and intuitive causal interpretation (see Corollary 7.3.2).

One key property of simple SCMs is that the solutions of the SCM always satisfy the conditional independencies implied by σ-separation (see Definition 6.2.7 and Corollary 6.2.13). 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. E.g., 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 (Forré and Mooij, 2019). Also, several causal discovery algorithms have already been proposed that work with simple SCMs, the first being an algorithm proposed by Forré and Mooij (2018). More recently, Mooij, Magliacane and Claassen (2019) showed that Local Causal Discovery (LCD) (Cooper, 1997), Y-structures (Mani, 2006) and the Joint Causal Inference framework (JCI) all apply to simple SCMs even though they were originally developed for acyclic SCMs only. Very recently it has been shown that even the well-known Fast Causal Inference (FCI) algorithm (Spirtes, Meek and Richardson, 1999; Zhang, 2008) is directly applicable to simple SCMs (Mooij and Claassen, 2020) and provides a consistent estimate of the Markov equivalence class (under the faithfulness assumption). 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 (i) reparametrizing and reducing the space of the exogenous variables of an SCM while preserving the causal and counterfactual semantics; (ii) extending and generalizing the identifiability results for (direct) causes and confounders; (iii) extending the graphs of SCMs to represent selection bias.

APPENDIX A: DIRECTED GRAPHICAL MODELS

A.1. Directed (mixed) graphs. In this section we introduce basic graphical notation and terminology for directed (mixed) graphs, where we do allow for cycles (Lauritzen, 1996; Richardson, 2003; Pearl, 2009; Forré and Mooij, 2017).

DEFINITION A.1.1 (Directed (mixed) graph).

1. *A* directed graph *is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of directed edges, which is a subset $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ of ordered pairs of nodes. Each element $(i, j) \in \mathcal{E}$ can be represented by the directed edge $i \to j$ or equivalently $j \leftarrow i$. In particular, $(i, i) \in \mathcal{E}$ represents a* self-cycle *$i \to i$.*
2. *A* directed mixed graph *is a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$, where the pair $(\mathcal{V}, \mathcal{E})$ forms a directed graph and $\mathcal{B}$ is a set of bidirected edges, which is a subset $\mathcal{B} \subseteq \{\{i, j\} : i, j \in \mathcal{V}, i \neq j\}$ of unordered (distinct) pairs of nodes. Each element $\{i, j\} \in \mathcal{B}$ can be represented by the bidirected edge $i \leftrightarrow j$ or equivalently $j \leftrightarrow i$. Note that a directed graph can be considered as a directed mixed graph without bidirected edges.*
3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph. A directed mixed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$ is a* subgraph *of $\mathcal{G}$ if $\tilde{\mathcal{V}} \subseteq \mathcal{V}$, $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ and $\tilde{\mathcal{B}} \subseteq \mathcal{B}$, in which case we write $\tilde{\mathcal{G}} \subseteq \mathcal{G}$. For a subset $\mathcal{W} \subseteq \mathcal{V}$, we define the* induced subgraph *of $\mathcal{G}$ on $\mathcal{W}$ by $\mathcal{G}_{\mathcal{W}} := (\mathcal{W}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$, where $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{B}}$ are the set of directed and bidirected edges in $\mathcal{E}$ and $\mathcal{B}$ respectively that lie in $\mathcal{W} \times \mathcal{W}$ and $\{\{i, j\} : i, j \in \mathcal{W}, i \neq j\}$ respectively.*
4. *A* walk *between $i, j \in \mathcal{V}$ in a directed mixed graph $\mathcal{G}$ is a tuple $(i_0, \epsilon_1, i_1, \epsilon_2, i_2, \ldots, \epsilon_n, i_n)$ of alternating nodes and edges in $\mathcal{G}$ for some $n \geq 0$, where all $i_0, \ldots, i_n \in \mathcal{V}$, all $\epsilon_1, \ldots, \epsilon_n \in \mathcal{E} \cup \mathcal{B}$*

such that $\epsilon_k \in \{i_{k-1} \to i_k, i_{k-1} \leftarrow i_k, i_{k-1} \leftrightarrow i_k\}$ *for all* $k = 1, \ldots, n$, *and it starts with node* $i_0 = i$ *and ends with node* $i_n = j$. *Note that* $n = 0$ *corresponds with a trivial walk consisting of a single node. If all nodes* $i_0, \ldots, i_n$ *are distinct, it is called a* path. *A walk (path) of the form* $i \to \cdots \to j$, *i.e.,* ϵ_k *is* $i_{k-1} \to i_k$ *for all* $k = 1, 2, \ldots, n$, *is called a* directed walk (path) *from* i *to* j.

5. *A* cycle *through* $i \in \mathcal{V}$ *in a directed mixed graph* $\mathcal{G}$ *is a directed path from* i *to some node* j *extended with the edge* $j \to i \in \mathcal{E}$. *In particular, a self-cycle* $i \to i \in \mathcal{E}$ *is a cycle. Note that a path cannot contain any cycles. A directed graph and a directed mixed graph are said to be* acyclic *if they contain no cycles, and are then referred to as a* directed acyclic graph (DAG) *and an* acyclic directed mixed graph (ADMG), *respectively.*
6. *For a directed (mixed) graph* $\mathcal{G}$ *and a node* $i \in \mathcal{V}$ *we define the set of* parents *of* i *by* $\mathrm{pa}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : j \to i \in \mathcal{E}\}$, *the set of children of* i *by* $\mathrm{ch}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : i \to j \in \mathcal{E}\}$, *the set of* ancestors *of* i *by*

$$\mathrm{an}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : \textit{there is a directed path from } j \textit{ to } i \textit{ in } \mathcal{G}\}$$

and the set of descendants *of* i *by*

$$\mathrm{de}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : \textit{there is a directed path from } i \textit{ to } j \textit{ in } \mathcal{G}\}\,.$$

Note that we have $\{i\} \cup \mathrm{pa}_{\mathcal{G}}(i) \subseteq \mathrm{an}_{\mathcal{G}}(i)$ *and* $\{i\} \cup \mathrm{ch}_{\mathcal{G}}(i) \subseteq \mathrm{de}_{\mathcal{G}}(i)$. *We can apply all these definitions to subsets* $\mathcal{U} \subseteq \mathcal{V}$ *by taking unions, for example* $\mathrm{pa}_{\mathcal{G}}(\mathcal{U}) := \cup_{i \in \mathcal{U}} \mathrm{pa}_{\mathcal{G}}(i)$. *A subset* $\mathcal{A} \subseteq \mathcal{V}$ *is called an* ancestral subset *in* $\mathcal{G}$ *if* $\mathcal{A} = \mathrm{an}_{\mathcal{G}}(\mathcal{A})$, *i.e.,* $\mathcal{A}$ *is closed under taking ancestors of* $\mathcal{A}$ *in* $\mathcal{G}$.
7. *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *be a directed mixed graph. We call* $\mathcal{G}$ strongly connected *if for every pair of distinct nodes* $i, j \in \mathcal{V}$, *the graph contains a cycle that passes through both* i *and* j. *The* strongly connected component of $i \in \mathcal{V}$, *denoted by* $\mathrm{sc}_{\mathcal{G}}(i)$, *is the maximal subset* $\mathcal{S} \subseteq \mathcal{V}$ *such that* $i \in \mathcal{S}$ *and the induced subgraph* $\mathcal{G}_{\mathcal{S}}$ *is strongly connected. Equivalently,* $\mathrm{sc}_{\mathcal{G}}(i) = \mathrm{an}_{\mathcal{G}}(i) \cap \mathrm{de}_{\mathcal{G}}(i)$.
8. *For a directed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, *we define the* graph of strongly connected components of $\mathcal{G}$ *as the directed graph* $\mathcal{G}^{\mathrm{sc}} := (\mathcal{V}^{\mathrm{sc}}, \mathcal{E}^{\mathrm{sc}})$, *where* $\mathcal{V}^{\mathrm{sc}}$ *are the strongly connected components of* $\mathcal{G}$, *i.e.,* $\mathcal{V}^{\mathrm{sc}}$ *are the equivalence classes in* $\mathcal{V}/\!\sim$ *with the equivalence relation* $i \sim j$ *iff* $i \in \mathrm{sc}_{\mathcal{G}}(j)$ *on* $\mathcal{V}$, *and* $\mathcal{E}^{\mathrm{sc}} = (\mathcal{E} \setminus \{i \to i : i \in \mathcal{V}\})/\!\sim$ *with the equivalence relation* $(i \to j) \sim (i' \to j')$ *iff* $i \sim i'$ *and* $j \sim j'$ *on* $\mathcal{E}$.

We omit the subscript $\mathcal{G}$ whenever it is clear which directed (mixed) graph $\mathcal{G}$ we are referring to.

LEMMA A.1.2 (DAG of strongly connected components). *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *be a directed graph. Then* $\mathcal{G}^{\mathrm{sc}}$, *the graph of strongly connected components of* $\mathcal{G}$, *is a DAG.*

A.2. The directed global Markov property. In this section we introduce the notation and terminology for the directed global Markov property. This directed global Markov property associates a conditional independence relation in the observational distribution of the SCM to each d-separation entailed by the graph. Here, we use a formulation of d-separation that generalizes d-separation for DAGs (Pearl, 1985) and m-separation for ADMGs (Richardson, 2003) and mDAGs (Evans, 2016), as discussed in more detail in (Forré and Mooij, 2017).

DEFINITION A.2.1 (Collider). *Let $\pi = (i_0, \epsilon_1, i_1, \epsilon_2, i_2, \ldots, \epsilon_n, i_n)$ be a walk (path) in a directed mixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$. A node i_k on π is called a* collider *on π if it is a non-endpoint node ($1 \leq k < n$) and the two edges $\epsilon_k, \epsilon_{k+1}$ meet head-to-head on i_k (i.e., if the subwalk $(i_{k-1}, \epsilon_k, i_k, \epsilon_{k+1}, i_{k+1})$ is of the form $i_{k-1} \to i_k \leftarrow i_{k+1}$, $i_{k-1} \leftrightarrow i_k \leftarrow i_{k+1}$, $i_{k-1} \to i_k \leftrightarrow i_{k+1}$, or $i_{k-1} \leftrightarrow i_k \leftrightarrow i_{k+1}$). The node i_k is called a* non-collider *on π otherwise, i.e., if it is an endpoint node ($k = 0$ or $k = n$) or if the subwalk $(i_{k-1}, \epsilon_k, i_k, \epsilon_{k+1}, i_{k+1})$ is of the form $i_{k-1} \to i_k \to i_{k+1}$, $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$, $i_{k-1} \leftarrow i_k \to i_{k+1}$, $i_{k-1} \leftrightarrow i_k \to i_{k+1}$, or $i_{k-1} \leftarrow i_k \leftrightarrow i_{k+1}$.*

Note in particular that the end points of a walk are non-colliders on the walk.

DEFINITION A.2.2 (d-separation). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and let $C \subseteq \mathcal{V}$ be a subset of nodes. A walk (path) $\pi = (i_0 \ldots i_n)$ in $\mathcal{G}$ is said to be* C-d-blocked or d-blocked by C *if*

1. *it contains a collider $i_k \notin \mathrm{an}_{\mathcal{G}}(C)$, or*
2. *it contains a non-collider $i_k \in C$.*

The walk (path) π is said to be C-d-open *if it is not d-blocked by C. For two subsets of nodes $A, B \subseteq \mathcal{V}$, we say that* A is d-separated from B given C in $\mathcal{G}$ *if all paths between a node in A and a node in B are d-blocked by C, and write*

$$A \underset{\mathcal{G}}{\overset{d}{\perp}} B \,|\, C\,.$$

The next lemma is a straightforward generalization of Lemma 3.3 in (Geiger, 1990) to the cyclic setting. It implies that it suffices to formulate d-separation in terms of paths rather than walks.

LEMMA A.2.3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph, $C \subseteq \mathcal{V}$ and $i, j \in \mathcal{V}$. There exists a C-d-open walk between i and j in $\mathcal{G}$ iff there exists a C-d-open path between i and j in $\mathcal{G}$.*

PROOF. It suffices to show that for every C-d-open walk between i and j in $\mathcal{G}$, there exists a C-d-open path between i and j in $\mathcal{G}$. Take a C-d-open walk $\pi = (i = i_0, \ldots, i_n = j)$. If a node ℓ occurs more than once in π, let i_j be the first occurrence of ℓ in π and i_k the last occurrence of ℓ in π. We now construct a new walk π' from π by removing the subwalk between i_j and i_k of π from π. It is easy to check that the new walk π' is still C-d-open. If ℓ is an endpoint on π', then i_j or i_k must be endpoint of π, and hence $\ell \notin C$. If ℓ is a non-endpoint non-collider on π', then also i_j or i_k must have been a non-endpoint non-collider on π, and hence $\ell \notin C$. If ℓ is a collider on π', then either (i) i_j or i_k are both colliders on π, and hence ℓ is ancestor of C in $\mathcal{G}$, or (ii) on the subwalk between i_j and i_k that was removed, there must be a directed path in $\mathcal{G}$ from i_j or i_k to a collider in $\mathrm{an}_{\mathcal{G}}(C)$, and hence, ℓ is in $\mathrm{an}_{\mathcal{G}}(C)$. The other nodes on π' cannot be responsible for C-d-blocking the walk, since they also occur (together with their adjacent edges) on π and they do not C-d-block π.

In π', the number of nodes that occur multiple times is at least one less than in π. Repeat this procedure until no repeated nodes are left. □

DEFINITION A.2.4 (Directed global Markov property). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathbb{P}_{\mathcal{V}}$ a probability distribution on $\boldsymbol{\mathcal{X}}_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$. The probability distribution $\mathbb{P}_{\mathcal{V}}$ obeys the* directed global Markov property *relative to* $\mathcal{G}$ *if for all subsets* $A, B, C \subseteq \mathcal{V}$ *we have*

$$A \underset{\mathcal{G}}{\overset{d}{\perp}} B \mid C \quad \Longrightarrow \quad \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{V}}}{\perp\!\!\!\perp} \boldsymbol{X}_B \mid \boldsymbol{X}_C \,,$$

i.e., $(X_i)_{i\in A}$ *and* $(X_i)_{i\in B}$ *are conditionally independent given* $(X_i)_{i\in C}$ *under* $\mathbb{P}_{\mathcal{V}}$*, where we take the canonical projections* $X_i : \boldsymbol{\mathcal{X}}_{\mathcal{V}} \to \mathcal{X}_i$ *as random variables.*

APPENDIX B: LINEAR SCMS

In this section we derive some results about (unique) solvability and marginalization for linear SCMs. Linear SCMs form a special class of SCMs that has seen much attention in the literature (see, e.g., Bollen, 1989; Hyttinen, Eberhardt and Hoyer, 2012).

Definition B.1 (Linear SCM). *We call an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$ linear *if each component of the causal mechanism is a linear combination of the endogenous and exogenous variables, that is*

$$f_i(\boldsymbol{x}, \boldsymbol{e}) = \sum_{j\in\mathcal{I}} B_{ij} x_j + \sum_{k\in\mathcal{J}} \Gamma_{ik} e_k \,,$$

where $i \in \mathcal{I}$, $B \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ *and* $\Gamma \in \mathbb{R}^{\mathcal{I}\times\mathcal{J}}$ *are matrices, and* $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$ *is a product probability measure*[20] *on* $\mathbb{R}^{\mathcal{J}}$*.*

For a subset $\mathcal{O} \subseteq \mathcal{I}$ we also use the shorthand vector-notation

$$\boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e}) = B_{\mathcal{O}\mathcal{I}} \boldsymbol{x} + \Gamma_{\mathcal{O}\mathcal{J}} \boldsymbol{e} \,.$$

A non-zero coefficient B_{ij} corresponds with a directed edge $i \to j$ in the augmented graph $\mathcal{G}^a(\mathcal{M})$, and a bidirected edge $i \leftrightarrow j$ in $\mathcal{G}(\mathcal{M})$ corresponds with a non-zero entry $(\Gamma\Gamma^T)_{ij}$ if the probability measure $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$ is a product of non-degenerate probability measures on $\mathbb{R}$.

For linear SCMs the solvability condition w.r.t. a subset, Definition 3.1.1, translates into a matrix condition. In order to state this condition we need to define the pseudoinverse (or the Moore-Penrose inverse) A^+ of a real matrix A (Penrose, 1955; Golub and Kahan, 1965). The *pseudoinverse of the matrix* A is defined by $A^+ := V\Sigma^+U^*$, where $A = U\Sigma V^*$ is the singular value decomposition of A and Σ^+ is obtained by replacing each non-zero entry on the diagonal of Σ by its reciprocal (Golub and Kahan, 1965). One of its useful properties is that $AA^+A = A$.

Proposition B.2 (Sufficient and necessary condition for solvability w.r.t. a subset for linear SCMs). *Let* $\mathcal{M}$ *be a linear SCM and* $\mathcal{L} \subseteq \mathcal{I}$ *and* $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$*. Then* $\mathcal{M}$ *is solvable w.r.t.* $\mathcal{L}$ *if and only if for the matrix* $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$*, for* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-almost every* $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ *and for all* $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ *the identity*

$$A_{\mathcal{L}\mathcal{L}} A^+_{\mathcal{L}\mathcal{L}} (B_{\mathcal{L}\mathcal{O}} \boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}} \boldsymbol{e}) = B_{\mathcal{L}\mathcal{O}} \boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}} \boldsymbol{e} \,,$$

where $A^+_{\mathcal{L}\mathcal{L}}$ *is the pseudoinverse of* $A_{\mathcal{L}\mathcal{L}}$*, is satisfied. Moreover, if* $\mathcal{M}$ *is solvable w.r.t.* $\mathcal{L}$*, then for every vector* $\boldsymbol{v} \in \mathbb{R}^{\mathcal{L}}$ *the mapping* $\boldsymbol{g}^{\boldsymbol{v}}_{\mathcal{L}} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{L}}$ *given by*

$$\boldsymbol{g}^{\boldsymbol{v}}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = A^+_{\mathcal{L}\mathcal{L}} (B_{\mathcal{L}\mathcal{O}} \boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}} \boldsymbol{e}) + [\mathbb{I}_{\mathcal{L}} - A^+_{\mathcal{L}\mathcal{L}} A_{\mathcal{L}\mathcal{L}}] \boldsymbol{v} \,,$$

is a measurable solution function for $\mathcal{M}$ *w.r.t.* $\mathcal{L}$*.*

[20] Note that we do not assume that the probability measure $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$ is Gaussian.

PROOF. Let $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$. For $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}$,

$$\begin{aligned}
&\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{e}) \\
&\iff \boldsymbol{x}_{\mathcal{L}} = B_{\mathcal{L}\mathcal{L}}\boldsymbol{x}_{\mathcal{L}} + B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e} \\
&\iff A_{\mathcal{L}\mathcal{L}}\boldsymbol{x}_{\mathcal{L}} = B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e} \\
&\iff \begin{cases} A_{\mathcal{L}\mathcal{L}}A^{+}_{\mathcal{L}\mathcal{L}}(B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e}) = B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e} \\ \exists_{\boldsymbol{v}\in\boldsymbol{\mathcal{X}}_{\mathcal{L}}} : \boldsymbol{x}_{\mathcal{L}} = A^{+}_{\mathcal{L}\mathcal{L}}(B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e}) + [\mathbb{I}_{\mathcal{L}} - A^{+}_{\mathcal{L}\mathcal{L}}A_{\mathcal{L}\mathcal{L}}]\boldsymbol{v}\,, \end{cases}
\end{aligned}$$

where the last equivalence follows from (Theorem 2, Penrose, 1955). □

For linear SCMs the unique solvability condition w.r.t. a subset translates into a matrix invertibility condition, as was already shown in (Hyttinen, Eberhardt and Hoyer, 2012).

PROPOSITION B.3 (Sufficient and necessary condition for unique solvability w.r.t. a subset for linear SCMs). *Let $\mathcal{M}$ be a linear SCM, $\mathcal{L} \subseteq \mathcal{I}$ and $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$. Then $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if the matrix $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. Moreover, if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then the mapping $\boldsymbol{g}_{\mathcal{L}} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{L}}$ given by*

$$\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = A^{-1}_{\mathcal{L}\mathcal{L}}(B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e})\,,$$

is a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{L}$.

PROOF. $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ iff for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ the linear system of equations

$$\begin{aligned}
&\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{e}) \\
&\iff \boldsymbol{x}_{\mathcal{L}} = B_{\mathcal{L}\mathcal{L}}\boldsymbol{x}_{\mathcal{L}} + B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e} \\
&\iff A_{\mathcal{L}\mathcal{L}}\boldsymbol{x}_{\mathcal{L}} = B_{\mathcal{L}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + \Gamma_{\mathcal{L}\mathcal{J}}\boldsymbol{e}
\end{aligned}$$

has a unique solution $\boldsymbol{x}_{\mathcal{L}} \in \mathcal{X}_{\mathcal{L}}$. Hence, $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ iff $A_{\mathcal{L}\mathcal{L}}$ is invertible. □

Note that if $A_{\mathcal{L}\mathcal{L}}$ is invertible, then $A^{+}_{\mathcal{L}\mathcal{L}} = A^{-1}_{\mathcal{L}\mathcal{L}}$ (see Lemma 1.3 in (Penrose, 1955)), and the matrix condition of Proposition B.2 is always satisfied and all the measurable solution functions $\boldsymbol{g}^{\boldsymbol{v}}_{\mathcal{L}}$ of Proposition B.2 are (up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set) equal to the solution function $\boldsymbol{g}_{\mathcal{L}}$ of Proposition B.3.

REMARK. *A sufficient condition for $A_{\mathcal{L}\mathcal{L}}$ to be invertible is that the spectral radius of $B_{\mathcal{L}\mathcal{L}}$ is less than one. If that is the case, then $A^{-1}_{\mathcal{L}\mathcal{L}} = \sum_{n=0}^{\infty}(B_{\mathcal{L}\mathcal{L}})^n$. Note that the non-zero coefficients of the matrix $B_{\mathcal{L}\mathcal{L}}$ represent the directed edges in the induced subgraph $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$. Similarly, for $n \in \mathbb{N}$, the coefficients of the matrix $(B_{\mathcal{L}\mathcal{L}})^n$ in the sum represent the sum of the product of the edge weights B_{ij} over directed paths of length n in the induced subgraph $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$.*

From Proposition 3.4.2 we know that an SCM is solvable w.r.t. $\mathcal{L}$ if and only if it is ancestrally solvable w.r.t. $\mathcal{L}$. In particular, this result also holds for linear SCMs. We saw in Example 3.4.3 that a similar result for unique solvability does not hold, that is, in general, it does not hold that unique solvability w.r.t. $\mathcal{L}$ implies ancestral unique solvability w.r.t. $\mathcal{L}$. For the class of linear SCMs we do have the following positive result.

PROPOSITION B.4 (Equivalent unique solvability conditions for linear SCMs). *For a linear SCM $\mathcal{M}$ and a subset $\mathcal{L} \subseteq \mathcal{I}$ the following are equivalent:*

1. *$\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$;*
2. *$\mathcal{M}$ is ancestrally uniquely solvable w.r.t. $\mathcal{L}$;*
3. *$\mathcal{M}$ is uniquely solvable w.r.t. each strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$.*

PROOF. It suffices to show (1) $\implies$ (2) and (1) $\iff$ (3). We start by showing that (1) $\implies$ (2). Let $\mathcal{V} \subseteq \mathcal{L}$ and denote $\mathcal{U} := \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{L}}}(\mathcal{V})$, then we need to show that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{U}$. From Proposition B.3 we know that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ iff the matrix $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. The matrix $A_{\mathcal{L}\mathcal{L}}$ is invertible iff the rows of $A_{\mathcal{L}\mathcal{L}}$ are all linearly independent. In particular, the rows of $A_{\mathcal{U}\mathcal{L}}$ are all linearly independent. Because $A_{\mathcal{U}\mathcal{L}} = [A_{\mathcal{U}\mathcal{U}}\, Z_{\mathcal{U}\mathcal{L}}]$, where $Z_{\mathcal{U}\mathcal{L}}$ is the zero matrix, we know that the rows of $A_{\mathcal{U}\mathcal{U}} = \mathbb{I}_{\mathcal{U}} - B_{\mathcal{U}\mathcal{U}}$ are also all linearly independent, and hence $A_{\mathcal{U}\mathcal{U}}$ is invertible.

Next, we show that (1) $\iff$ (3). Observe that the strongly connected components of $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$ form a partition of the set $\mathcal{L}$ and that the directed mixed graph $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$ and the directed graph $\mathcal{G}^a(\mathcal{M})_{\mathcal{L}}$ have the same strongly connected components. Because, by Lemma A.1.2, the graph of strongly connected components $\mathcal{G}^{\mathrm{sc}}$ of the directed graph $\mathcal{G}^a(\mathcal{M})_{\mathcal{L}}$ is a DAG, the square matrix $B_{\mathcal{L}\mathcal{L}}$ can be permuted to an upper triangular block matrix $\tilde{B}_{\mathcal{L}\mathcal{L}}$, where for each diagonal block $\tilde{B}_{\mathcal{V}\mathcal{V}}$ of $\tilde{B}_{\mathcal{L}\mathcal{L}}$ the set of nodes $\mathcal{V}$ is a strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$.

Without loss of generality we assume now that $B_{\mathcal{L}\mathcal{L}}$ is an upper triangular block matrix. From Proposition B.3 it follows that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ iff the matrix $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. Because $B_{\mathcal{L}\mathcal{L}}$ is an upper triangular block matrix, we know that $A_{\mathcal{L}\mathcal{L}}$ is an upper triangular block matrix, where for each diagonal block $A_{\mathcal{V}\mathcal{V}}$ of $A_{\mathcal{L}\mathcal{L}}$ the set of nodes $\mathcal{V}$ is a strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$. Since an upper triangular block matrix $A_{\mathcal{L}\mathcal{L}}$ is invertible iff every diagonal block in $A_{\mathcal{L}\mathcal{L}}$ is invertible, we have that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ iff $\mathcal{M}$ is uniquely solvable w.r.t. each strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$. □

Under the condition of unique solvability w.r.t. a subset $\mathcal{L}$ we can define the marginalization w.r.t. $\mathcal{L}$ of a linear SCM by mere substitution.

PROPOSITION B.5 (Marginalization of a linear SCM). *Given a linear SCM $\mathcal{M}$ and a subset $\mathcal{L} \subsetneq \mathcal{I}$ of endogenous variables such that $\mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. Then there exists a marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ that is linear and with marginal causal mechanism $\tilde{\boldsymbol{f}} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{O}}$ given by*

$$\tilde{\boldsymbol{f}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = [B_{\mathcal{O}\mathcal{O}} + B_{\mathcal{O}\mathcal{L}} A_{\mathcal{L}\mathcal{L}}^{-1} B_{\mathcal{L}\mathcal{O}}]\boldsymbol{x}_{\mathcal{O}} + [B_{\mathcal{O}\mathcal{L}} A_{\mathcal{L}\mathcal{L}}^{-1} \Gamma_{\mathcal{L}\mathcal{J}} + \Gamma_{\mathcal{O}\mathcal{J}}]\boldsymbol{e}\,,$$

where $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$. Moreover, this marginalization respects the latent projection, i.e., $\big(\mathcal{G}^a \circ \mathrm{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\mathrm{marg}(\mathcal{L}) \circ \mathcal{G}^a\big)(\mathcal{M})$.

PROOF. By the definition of marginalization and Proposition B.3 the marginal causal mechanism $\tilde{\boldsymbol{f}}$ is given by

$$\begin{aligned}
\tilde{\boldsymbol{f}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) &:= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{e}) \\
&= B_{\mathcal{O}\mathcal{O}}\boldsymbol{x}_{\mathcal{O}} + B_{\mathcal{O}\mathcal{L}}\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) + \Gamma_{\mathcal{O}\mathcal{J}}\boldsymbol{e} \\
&= [B_{\mathcal{O}\mathcal{O}} + B_{\mathcal{O}\mathcal{L}} A_{\mathcal{L}\mathcal{L}}^{-1} B_{\mathcal{L}\mathcal{O}}]\boldsymbol{x}_{\mathcal{O}} + [B_{\mathcal{O}\mathcal{L}} A_{\mathcal{L}\mathcal{L}}^{-1} \Gamma_{\mathcal{L}\mathcal{J}} + \Gamma_{\mathcal{O}\mathcal{J}}]\boldsymbol{e}\,.
\end{aligned}$$

From Proposition B.4 and 5.2.6 it follows that the marginalization respects the latent projection. □

From Theorem 5.1.7 we know that $\mathcal{M}$ and its marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ over $\mathcal{L}$ are observationally, interventionally and counterfactually equivalent w.r.t. $\mathcal{O}$. A similar result can also be found in (Hyttinen, Eberhardt and Hoyer, 2012). In contrast to non-linear SCMs, this class of linear SCMs has the convenient property that every marginalization of a model of this class respects the latent projection. Moreover, the subclass of simple linear SCMs is even closed under marginalization.

APPENDIX C: PROOFS

This section contains the proofs of all the theoretical results in the main text. Some of the proofs will rely on the measure theoretic terminology and results of Appendix D.

PROOF OF PROPOSITION 2.3.3. Let $i \in \mathcal{I}$. Note that Definition 2.2.1 can alternatively be formulated as follows: for $k \in \mathcal{I} \cup \mathcal{J}$, $k \notin \mathrm{pa}(i)$ if and only if there exists a measurable mapping $\hat{f}_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

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

and either $k \in \mathcal{I}$ and there exists $\hat{x}_k \in \mathcal{X}_k$ such that $\hat{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \hat{f}_i(\boldsymbol{x}_{\setminus k}, \hat{x}_k, \boldsymbol{e})$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, or $k \in \mathcal{J}$ and there exists $\hat{e}_k \in \mathcal{E}_k$ such that $\hat{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \hat{f}_i(\boldsymbol{x}, \boldsymbol{e}_{\setminus k}, \hat{e}_k)$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}}$. By repeatedly applying (this formulation of) Definition 2.2.1 to all $k \notin \mathrm{pa}(i)$, we obtain the existence of a measurable mapping $\tilde{f}_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$ and $\hat{\boldsymbol{x}}_{\setminus \mathrm{pa}(i)} \in \boldsymbol{\mathcal{X}}_{\setminus \mathrm{pa}(i)}$, $\hat{\boldsymbol{e}}_{\setminus \mathrm{pa}(i)} \in \boldsymbol{\mathcal{E}}_{\setminus \mathrm{pa}(i)}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

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

and for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

$$\tilde{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \tilde{f}_i(\boldsymbol{x}_{\mathrm{pa}(i)}, \hat{\boldsymbol{x}}_{\setminus \mathrm{pa}(i)}, \boldsymbol{e}_{\mathrm{pa}(i)}, \hat{\boldsymbol{e}}_{\setminus \mathrm{pa}(i)}).$$

Define the SCM $\tilde{\mathcal{M}}$ as $\mathcal{M}$ except that its causal mechanism is $\tilde{\boldsymbol{f}}$ instead of $\boldsymbol{f}$. Then $\tilde{\mathcal{M}}$ is structurally minimal and equivalent to $\mathcal{M}$. □

PROOF OF PROPOSITION 2.4.3. The $\mathrm{do}(I, \boldsymbol{\xi}_I)$ operation on $\mathcal{M}$ completely removes the functional dependence on $\boldsymbol{x}$ and $\boldsymbol{e}$ from the f_i components for $i \in I$ and hence the corresponding incoming directed and bidirected edges on nodes in I from the (augmented) graph. □

PROOF OF PROPOSITION 2.4.4. The first statement follows directly from Definitions 2.4.1 and 2.4.2. For the second statement, note that a perfect intervention can only remove parental relations, and therefore will never introduce a cycle. □

PROOF OF PROPOSITION 2.5.2. This follows directly from Definitions 2.4.1 and 2.5.1. □

PROOF OF PROPOSITION 2.5.4. This follows directly from Definition 2.5.1 and 2.5.3. □

PROOF OF PROPOSITION 2.5.5. Applying the intervention $\mathrm{do}(I)$ on the graph $\mathcal{G}$ removes all the incoming edges from the nodes in I. Now, if we perform the twin operation w.r.t. K on this graph $\mathrm{do}(I)(\mathcal{G})$, then we copy the same edges as if we had twinned the graph $\mathcal{G}$ w.r.t. K, except those edges that do point to one of the nodes in I. Hence, if we apply the intervention $\mathrm{do}(I \cup I')$ on the graph $\mathrm{twin}(K)(\mathcal{G})$, which removes all incoming edges of both I and its copy I', then we clearly obtain the same graph. □

Proof of Proposition 2.5.6. The additional edges introduced by the twinning operation cannot lead to a directed cycle involving both copied and original nodes, because there are no edges pointing from copied nodes to original nodes (i.e., of the form $i' \to v$ with $i' \in I'$ and $v \in \mathcal{V}$). Directed cycles involving only original nodes are absent by assumption, and directed cycles involving only copied nodes as well since they would correspond with a directed cycle in the original directed graph. $\square$

Proof of Theorem 3.1.3. First we define the solution space $\boldsymbol{\mathcal{S}}(\mathcal{M})$ of $\mathcal{M}$ by

$$\boldsymbol{\mathcal{S}}(\mathcal{M}) := \{(\boldsymbol{e}, \boldsymbol{x}) \in \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} : \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})\}\,.$$

This is a measurable set, since $\boldsymbol{\mathcal{S}}(\mathcal{M}) = \boldsymbol{h}^{-1}(\Delta)$, where $\boldsymbol{h} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}$ is the measurable mapping defined by $\boldsymbol{h}(\boldsymbol{e}, \boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}))$ and Δ is the set defined by $\{(\boldsymbol{x}, \boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{\mathcal{X}}\}$, which is measurable since $\boldsymbol{\mathcal{X}}$ is Hausdorff. Note that

$$\boldsymbol{\mathcal{A}} := \boldsymbol{pr}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}}(\mathcal{M})) = \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \exists \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \text{ s.t. } \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})\}\,,$$

is an analytic set because the projection $\boldsymbol{pr}_{\boldsymbol{\mathcal{E}}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}$ is a measurable mapping between standard measurable spaces (Lemma D.3).

Suppose that (1) holds, that is, $\mathcal{M}$ has a solution. Then there exists a pair of random variables $(\boldsymbol{E}, \boldsymbol{X}) : \Omega \to \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$ $\mathbb{P}$-a.s.. Note that

$$\begin{aligned}\{\omega \in \Omega : \boldsymbol{X}(\omega) = \boldsymbol{f}\big(\boldsymbol{X}(\omega), \boldsymbol{E}(\omega)\big)\} &\subseteq \{\omega \in \Omega : \exists \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \text{ s.t. } \boldsymbol{x} = \boldsymbol{f}\big(\boldsymbol{x}, \boldsymbol{E}(\omega)\big)\} \\ &\subseteq \boldsymbol{E}^{-1}\Big(\{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \exists \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \text{ s.t. } \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})\}\Big) \\ &= \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{A}}).\end{aligned}$$

By Lemma D.6, $\boldsymbol{\mathcal{A}}$ is $\mathbb{P}^{\boldsymbol{E}}$-measurable because it is analytic, and we can write $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{B}} \dot{\cup} \boldsymbol{\mathcal{N}}$ with $\boldsymbol{\mathcal{B}} \subseteq \boldsymbol{\mathcal{E}}$ measurable and $\boldsymbol{\mathcal{N}}$ a $\mathbb{P}^{\boldsymbol{E}}$-null set. Hence $\boldsymbol{E}^{-1}(\boldsymbol{\mathcal{A}}) = \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{B}}) \cup \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{N}})$ where $\boldsymbol{E}^{-1}(\boldsymbol{\mathcal{N}})$ is a $\mathbb{P}$-null set. Therefore,

$$\boldsymbol{E}^{-1}(\boldsymbol{\mathcal{B}}) \supseteq \{\omega \in \Omega : \boldsymbol{X}(\omega) = \boldsymbol{f}\big(\boldsymbol{X}(\omega), \boldsymbol{E}(\omega)\big)\} \setminus \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{N}})$$

which implies that $\mathbb{P}(\boldsymbol{E}^{-1}(\boldsymbol{\mathcal{B}})) = 1$. Hence, $\boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{A}}$ is a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. In other words, 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 $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, i.e., (2) holds.

Suppose that (2) holds. Then $\boldsymbol{\mathcal{E}} \setminus \boldsymbol{pr}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}}(\mathcal{M}))$ is a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. By application of the measurable selection theorem D.8, there exists a measurable $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, $\boldsymbol{g}(\boldsymbol{e}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{e}), \boldsymbol{e})$. Hence (3) holds.

Suppose (3) holds. Let $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be the causal mechanism of a structurally minimal SCM that is equivalent to $\mathcal{M}$ (see Proposition 2.3.3). In particular, for any $\boldsymbol{\epsilon}_{\setminus \mathrm{pa}(\mathcal{I})} \in \boldsymbol{\mathcal{E}}_{\setminus \mathrm{pa}(\mathcal{I})}$, we have that $\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}, \boldsymbol{\epsilon}_{\setminus \mathrm{pa}(\mathcal{I})})$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$. This means that we may also consider $\tilde{\boldsymbol{f}}$ as a mapping $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$. By applying Lemma D.10 to the canonical projection $\boldsymbol{pr}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}$ and using the equivalence of $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$, we obtain that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}}$-almost all $\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}$ there exists $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ with $\boldsymbol{x} = \tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{I})})$. By applying the implication (2) $\implies$ (3) to $\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}$ and $\tilde{\boldsymbol{f}}$, we conclude the existence of a measurable $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}}$-almost all $\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})}$, $\boldsymbol{g}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}) = \tilde{\boldsymbol{f}}(\boldsymbol{g}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}), \boldsymbol{e}_{\mathrm{pa}(\mathcal{I})})$. Once more using Lemma D.10, we obtain that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, $\boldsymbol{g}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}), \boldsymbol{e})$. In other words, (4) holds.

Lastly, suppose that (4) holds, that is there exists a measurable solution function $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$. Then the measurable mappings $\boldsymbol{E} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}$ and $\boldsymbol{X} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$, defined by $\boldsymbol{E}(\boldsymbol{e}) := \boldsymbol{e}$ and $\boldsymbol{X}(\boldsymbol{e}) := \boldsymbol{g}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})})$ respectively, define a pair of random variables $(\boldsymbol{X}, \boldsymbol{E})$ such that $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$ holds a.s. and hence $(\boldsymbol{X}, \boldsymbol{E})$ is a solution. Hence (1) holds. □

PROOF OF PROPOSITION 3.1.5. Let $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be the causal mechanism of a structurally minimal SCM that is equivalent to $\mathcal{M}$ (see Proposition 2.3.3). In particular, for any $\boldsymbol{\epsilon}_{\setminus \mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{E}}_{\setminus \mathrm{pa}(\mathcal{O})}$ and $\boldsymbol{\xi}_{\setminus \mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{X}}_{\setminus \mathrm{pa}(\mathcal{O})}$, we have that for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, $\tilde{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{\xi}_{\setminus \mathrm{pa}(\mathcal{O})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{\epsilon}_{\setminus \mathrm{pa}(\mathcal{O})})$. This means that we may also consider $\tilde{\boldsymbol{f}}$ as a mapping $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}$.

Consider the set

$$\tilde{\boldsymbol{\mathcal{S}}} := \{(\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{x}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}}, \boldsymbol{x}_{\mathcal{O}}) \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \times \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}} \,:\, \boldsymbol{x}_{\mathcal{O}} = \tilde{\boldsymbol{f}}_{\mathcal{O}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})})\}\,.$$

By similar reasoning as in the proof of Theorem 3.1.3, $\tilde{\boldsymbol{\mathcal{S}}}$ is measurable.

By assumption, 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 space $\{\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} : \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})\}$ is non-empty and σ-compact. By applying Lemma D.10 to the canonical projection $\boldsymbol{pr}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}$ and using the equivalence of $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$, we obtain that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}}$-almost every $\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}$ and for all $\boldsymbol{x}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}}$ the space

$$\tilde{\boldsymbol{\mathcal{S}}}_{(\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{x}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}})} := \{\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} : \boldsymbol{x}_{\mathcal{O}} = \tilde{\boldsymbol{f}}_{\mathcal{O}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})})\}$$

is non-empty and σ-compact.

The second measurable selection theorem, Theorem D.9, now implies that there exists a measurable $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}}$-almost every $\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}$ and for all $\boldsymbol{x}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}}$

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

Once more applying Lemma D.10, we obtain that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) \implies \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e}).$$

Hence $\mathcal{M}$ is solvable w.r.t. $\mathcal{O}$. □

PROOF OF PROPOSITION 3.2.2. Let $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be the causal mechanism of a structurally minimal SCM $\tilde{\mathcal{M}}$ that is equivalent to $\mathcal{M}$ (see Proposition 2.3.3). For a subset $\mathcal{O} \subseteq \mathcal{I}$ consider the induced subgraph $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ of the augmented graph $\mathcal{G}^a(\mathcal{M})$ on $\mathcal{O}$. Then the acyclicity of $\mathcal{G}^a(\mathcal{M})$ implies that the induced subgraph $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ is acyclic, and hence there exists a topological ordering on the nodes $\mathcal{O}$. We can substitute the components $\tilde{f}_i$ of the causal mechanism $\tilde{\boldsymbol{f}}$ for $i \in \mathcal{O}$ into each other along this topological ordering. This gives a measurable solution function $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ for $\tilde{\mathcal{M}}$, and hence for $\mathcal{M}$. It is clear from the acyclic structure that this mapping $\boldsymbol{g}_{\mathcal{O}}$ is independent of the choice of the topological ordering and is the only solution function for $\mathcal{M}$. Therefore, $\tilde{\mathcal{M}}$ is uniquely solvable w.r.t. $\mathcal{O}$, and so is $\mathcal{M}$. □

PROOF OF PROPOSITION 3.2.4. This follows immediately from Definition 2.2.2 and 3.2.1. □

PROOF OF THEOREM 3.2.5. Suppose that (1) holds. By Proposition 3.1.5 there exists a measurable solution function $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ for $\mathcal{M}$ w.r.t. $\mathcal{O}$. Then for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\setminus\mathcal{O}}$ we have that $\boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})})$ is a solution of $\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})$. Hence, because of (1), for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\setminus\mathcal{O}}$ we have that $\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})$ implies $\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})})$. Thus, $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{O}$, that is, (2) holds.

Suppose that (2) holds. Let $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{O}$. Then, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

This implies (1).

For the last statement, assume that $\mathcal{M}$ is uniquely solvable. Let $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ be a measurable solution function. Then there exists a measurable set $\boldsymbol{B} \subseteq \boldsymbol{\mathcal{E}}$ with $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{B}) = 1$ and for all $\boldsymbol{e} \in \boldsymbol{B}$,

$$\forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}} : \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) \implies \boldsymbol{x} = \boldsymbol{g}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}).$$

The existence of a solution for $\mathcal{M}$ follows directly from Theorem 3.1.3. Each solution $(\boldsymbol{X}, \boldsymbol{E}) : \Omega \to \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}}$ of $\mathcal{M}$ satisfies $\boldsymbol{X}(\omega) = \boldsymbol{f}(\boldsymbol{X}(\omega), \boldsymbol{E}(\omega))$ $\mathbb{P}$-a.s.. In addition, it satisfies $\boldsymbol{E}(\omega) \in \boldsymbol{B}$ $\mathbb{P}$-a.s., since $\mathbb{P} \circ \boldsymbol{E}^{-1} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$. Hence, it satisfies $\boldsymbol{X}(\omega) = \boldsymbol{g}(\boldsymbol{E}(\omega)_{\mathrm{pa}(\mathcal{I})})$ $\mathbb{P}$-a.s.. Thus for every solution $(\boldsymbol{X}, \boldsymbol{E})$ the associated observational distribution is the push-forward of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ under $\boldsymbol{g} \circ \boldsymbol{pr}_{\mathrm{pa}(\mathcal{I})}$. □

PROOF OF PROPOSITION 3.3.2. Let $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{O}$. Then the mapping $\tilde{\boldsymbol{g}}_{\mathrm{pa}(\mathcal{O})\cup\mathcal{O}} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\cup\mathcal{O}}$ defined by $\tilde{\boldsymbol{g}}_{\mathrm{pa}(\mathcal{O})\cup\mathcal{O}}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) := (\boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}))$ is a measurable solution function for $\mathcal{M}_{\mathrm{do}(\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}, \boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}})}$ w.r.t. $\mathrm{pa}(\mathcal{O}) \cup \mathcal{O}$. Moreover, the mapping $\hat{\boldsymbol{g}} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ defined by $\hat{\boldsymbol{g}}(\boldsymbol{e}_{\mathrm{pa}(\mathcal{I})}) := (\boldsymbol{\xi}_{\mathcal{I}\setminus\mathcal{O}}, \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}))$ is a measurable solution function for $\mathcal{M}_{\mathrm{do}(\mathcal{I}\setminus\mathcal{O}, \boldsymbol{\xi}_{\mathcal{I}\setminus\mathcal{O}})}$. If $\mathcal{M}$ w.r.t. $\mathcal{O}$ is uniquely solvable w.r.t. $\mathcal{O}$, it follows that also $\mathcal{M}_{\mathrm{do}(\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}, \boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}})}$ is uniquely solvable w.r.t. $\mathcal{O}$ and that $\mathcal{M}_{\mathrm{do}(\mathcal{I}\setminus\mathcal{O}, \boldsymbol{\xi}_{\mathcal{I}\setminus\mathcal{O}})}$ is uniquely solvable. □

PROOF OF PROPOSITION 3.4.2. It suffices to show that solvability of $\mathcal{M}$ w.r.t. $\mathcal{O}$ implies ancestral solvability w.r.t. $\mathcal{O}$. Solvability of $\mathcal{M}$ w.r.t. $\mathcal{O}$ implies that there exists a measurable mapping $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

Let $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be the causal mechanism of a structurally minimal SCM $\tilde{\mathcal{M}}$ that is equivalent to $\mathcal{M}$ (see Proposition 2.3.3). Let $\mathcal{P} := \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$ for some $\mathcal{A} \subseteq \mathcal{O}$. Then for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\begin{cases} \boldsymbol{x}_{\mathcal{P}} & = (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) \\ \boldsymbol{x}_{\mathcal{O}\setminus\mathcal{P}} & = (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\setminus\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) \end{cases} \Longrightarrow \begin{cases} \boldsymbol{x}_{\mathcal{P}} & = \tilde{\boldsymbol{f}}_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{P})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{P})}) \\ \boldsymbol{x}_{\mathcal{O}\setminus\mathcal{P}} & = \tilde{\boldsymbol{f}}_{\mathcal{O}\setminus\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O}\setminus\mathcal{P})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O}\setminus\mathcal{P})}) \,. \end{cases}$$

Since $\mathrm{pa}(\mathcal{P}) \setminus \mathcal{P} \subseteq \mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}$, we have that in particular for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

This implies that the mapping $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{P}}$ cannot depend on elements different from $\mathrm{pa}(\mathcal{P})$. Moreover, it follows from the definition of $\mathcal{P}$ that $(\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathrm{pa}(\mathcal{P}) = \mathrm{pa}(\mathcal{P}) \setminus \mathcal{P}$ and thus we have $\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O} =$

$(\mathrm{pa}(\mathcal{P}) \setminus \mathcal{P}) \cup (\mathrm{pa}(\mathcal{O}) \setminus (\mathcal{O} \cup \mathrm{pa}(\mathcal{P})))$. Now, pick an element $\hat{\boldsymbol{x}}_{\mathrm{pa}(\mathcal{O})\setminus(\mathcal{O}\cup\mathrm{pa}(\mathcal{P}))} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus(\mathcal{O}\cup\mathrm{pa}(\mathcal{P}))}$ and define the mapping $\tilde{\boldsymbol{g}}_{\mathcal{P}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{P})\setminus\mathcal{P}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{P})} \to \boldsymbol{\mathcal{X}}_{\mathcal{P}}$ by

$$\tilde{\boldsymbol{g}}_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{P})\setminus\mathcal{P}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{P})}) := (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{P})\setminus\mathcal{P}}, \hat{\boldsymbol{x}}_{\mathrm{pa}(\mathcal{O})\setminus(\mathcal{O}\cup\mathrm{pa}(\mathcal{P}))}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) .$$

Then, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{P}} = \tilde{\boldsymbol{g}}_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{P})\setminus\mathcal{P}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{P})}) \quad \iff \quad \boldsymbol{x}_{\mathcal{P}} = (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) .$$

Together this gives that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{P}} = \tilde{\boldsymbol{g}}_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{P})\setminus\mathcal{P}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{P})}) \quad \implies \quad \boldsymbol{x}_{\mathcal{P}} = \tilde{\boldsymbol{f}}_{\mathcal{P}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{P})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{P})}) .$$

which is equivalent to the statement that $\mathcal{M}$ is solvable w.r.t. $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$. □

LEMMA C.1. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. two subsets $A, B \subseteq \mathcal{I}$ that satisfy $A \subseteq B$ and $\mathrm{pa}(A) \setminus A \subseteq \mathrm{pa}(B) \setminus B$. Let $\boldsymbol{g}_A : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(A)\setminus A} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(A)} \to \boldsymbol{\mathcal{X}}_A$ and $\boldsymbol{g}_B : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(B)\setminus B} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(B)} \to \boldsymbol{\mathcal{X}}_B$ be measurable solution functions for $\mathcal{M}$ w.r.t. A and B, respectively. Then for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*

$$\boldsymbol{g}_A(\boldsymbol{x}_{\mathrm{pa}(A)\setminus A}, \boldsymbol{e}_{\mathrm{pa}(A)}) = (\boldsymbol{g}_B)_A(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) .$$

PROOF. Without loss of generality, we assume that $\mathcal{M}$ is structurally minimal (see Proposition 2.3.3). Let $\bar{\boldsymbol{\mathcal{E}}} \subseteq \boldsymbol{\mathcal{E}}$ be a measurable set with $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\bar{\boldsymbol{\mathcal{E}}}) = 1$ such that for all $\boldsymbol{e} \in \bar{\boldsymbol{\mathcal{E}}}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$:

$$\boldsymbol{x}_A = \boldsymbol{g}_A(\boldsymbol{x}_{\mathrm{pa}(A)\setminus A}, \boldsymbol{e}_{\mathrm{pa}(A)}) \iff \boldsymbol{x}_A = \boldsymbol{f}_A(\boldsymbol{x}_{\mathrm{pa}(A)}, \boldsymbol{e}_{\mathrm{pa}(A)})$$

and

$$\boldsymbol{x}_B = \boldsymbol{g}_B(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) \iff \boldsymbol{x}_B = \boldsymbol{f}_B(\boldsymbol{x}_{\mathrm{pa}(B)}, \boldsymbol{e}_{\mathrm{pa}(B)}) .$$

Now let $\boldsymbol{e} \in \bar{\boldsymbol{\mathcal{E}}}$ and let $\boldsymbol{x}_{A\cup\mathrm{pa}(B)\setminus B} \in \boldsymbol{\mathcal{X}}_{A\cup\mathrm{pa}(B)\setminus B}$. Then

$$\begin{aligned}
&\boldsymbol{x}_A = (\boldsymbol{g}_B)_A(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) \\
&\implies \begin{cases} & \boldsymbol{x}_A = (\boldsymbol{g}_B)_A(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) \\ \exists \boldsymbol{x}_{B\setminus A} \in \boldsymbol{\mathcal{X}}_{B\setminus A} : & \boldsymbol{x}_{B\setminus A} = (\boldsymbol{g}_B)_{B\setminus A}(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) \end{cases} \\
&\implies \exists \boldsymbol{x}_{B\setminus A} \in \boldsymbol{\mathcal{X}}_{B\setminus A} : \quad \boldsymbol{x}_B = \boldsymbol{g}_B(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) \\
&\implies \exists \boldsymbol{x}_{B\setminus A} \in \boldsymbol{\mathcal{X}}_{B\setminus A} : \quad \boldsymbol{x}_B = \boldsymbol{f}_B(\boldsymbol{x}_{\mathrm{pa}(B)}, \boldsymbol{e}_{\mathrm{pa}(B)}) \\
&\implies \exists \boldsymbol{x}_{B\setminus A} \in \boldsymbol{\mathcal{X}}_{B\setminus A} : \quad \boldsymbol{x}_A = \boldsymbol{f}_A(\boldsymbol{x}_{\mathrm{pa}(A)}, \boldsymbol{e}_{\mathrm{pa}(A)}) \\
&\implies \boldsymbol{x}_A = \boldsymbol{f}_A(\boldsymbol{x}_{\mathrm{pa}(A)}, \boldsymbol{e}_{\mathrm{pa}(A)}) \\
&\implies \boldsymbol{x}_A = \boldsymbol{g}_A(\boldsymbol{x}_{\mathrm{pa}(A)\setminus A}, \boldsymbol{e}_{\mathrm{pa}(A)}) ,
\end{aligned}$$

where the exists-quantifier could be omitted because the expression it binds to does not depend on $\boldsymbol{x}_{B\setminus A}$ (from the assumptions it follows that $(A \cup \mathrm{pa}(A)) \cap (B \setminus A) = \emptyset$). Hence, for all $\boldsymbol{e} \in \bar{\boldsymbol{\mathcal{E}}}$ and all $\boldsymbol{x}_{A\cup\mathrm{pa}(B)\setminus B} \in \boldsymbol{\mathcal{X}}_{A\cup\mathrm{pa}(B)\setminus B}$

$$\boldsymbol{x}_A = (\boldsymbol{g}_B)_A(\boldsymbol{x}_{\mathrm{pa}(B)\setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) \implies \boldsymbol{x}_A = \boldsymbol{g}_A(\boldsymbol{x}_{\mathrm{pa}(A)\setminus A}, \boldsymbol{e}_{\mathrm{pa}(A)}) .$$

Hence, for all $\boldsymbol{e} \in \bar{\boldsymbol{\mathcal{E}}}$ and all $\boldsymbol{x}_{A \cup \mathrm{pa}(B) \setminus B} \in \boldsymbol{\mathcal{X}}_{A \cup \mathrm{pa}(B) \setminus B}$

$$(\boldsymbol{g}_B)_A(\boldsymbol{x}_{\mathrm{pa}(B) \setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) = \boldsymbol{g}_A(\boldsymbol{x}_{\mathrm{pa}(A) \setminus A}, \boldsymbol{e}_{\mathrm{pa}(A)}) .$$

Since this expression does not depend on $\boldsymbol{x}_{(B \setminus A) \cup \mathcal{I} \setminus (B \cup \mathrm{pa}(B))}$, we conclude from Lemma D.11.(2) that for all $\boldsymbol{e} \in \bar{\boldsymbol{\mathcal{E}}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$(\boldsymbol{g}_B)_A(\boldsymbol{x}_{\mathrm{pa}(B) \setminus B}, \boldsymbol{e}_{\mathrm{pa}(B)}) = \boldsymbol{g}_A(\boldsymbol{x}_{\mathrm{pa}(A) \setminus A}, \boldsymbol{e}_{\mathrm{pa}(A)}) .$$

□

PROOF OF PROPOSITION 3.4.4. Without loss of generality, we assume that $\mathcal{M}$ is structurally minimal (see Proposition 2.3.3). Define $\mathcal{C} := \mathcal{A} \cap \tilde{\mathcal{A}}$ and $\mathcal{D} := \mathcal{A} \cup \tilde{\mathcal{A}}$. Let $\boldsymbol{g}_{\mathcal{A}}$, $\boldsymbol{g}_{\tilde{\mathcal{A}}}$ be measurable solution functions for $\mathcal{M}$ w.r.t. $\mathcal{A}$ and $\tilde{\mathcal{A}}$, respectively. Note that $\mathrm{pa}(\mathcal{C}) \setminus \mathcal{C} \subseteq \mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}$ and similarly $\mathrm{pa}(\mathcal{C}) \setminus \mathcal{C} \subseteq \mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}$. Indeed, for $c \in \mathrm{pa}(\mathcal{C})$: if $c \in \mathcal{O}$ then $c \in \mathcal{C}$ because $\mathcal{A}$ and $\tilde{\mathcal{A}}$ are both ancestral in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$, while if $c \notin \mathcal{O}$ then $c \notin \mathcal{A}$ and $c \notin \tilde{\mathcal{A}}$. Hence by Lemma C.1, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$(\boldsymbol{g}_{\mathcal{A}})_{\mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{A})}) = (\boldsymbol{g}_{\tilde{\mathcal{A}}})_{\mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}}, \boldsymbol{e}_{\mathrm{pa}(\tilde{\mathcal{A}})}) .$$

Hence for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\begin{aligned}
& \boldsymbol{x}_{\mathcal{D}} = \boldsymbol{f}_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{e}) \\
\iff & \begin{cases} \boldsymbol{x}_{\mathcal{A} \setminus \mathcal{C}} & = \boldsymbol{f}_{\mathcal{A} \setminus \mathcal{C}}(\boldsymbol{x}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{C}} & = \boldsymbol{f}_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{C}} & = \boldsymbol{f}_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{e}) \\ \boldsymbol{x}_{\tilde{\mathcal{A}} \setminus \mathcal{C}} & = \boldsymbol{f}_{\tilde{\mathcal{A}} \setminus \mathcal{C}}(\boldsymbol{x}, \boldsymbol{e}) \end{cases} \\
\iff & \begin{cases} \boldsymbol{x}_{\mathcal{A} \setminus \mathcal{C}} & = (\boldsymbol{g}_{\mathcal{A}})_{\mathcal{A} \setminus \mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{A})}) \\ \boldsymbol{x}_{\mathcal{C}} & = (\boldsymbol{g}_{\mathcal{A}})_{\mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{A})}) \\ \boldsymbol{x}_{\mathcal{C}} & = (\boldsymbol{g}_{\tilde{\mathcal{A}}})_{\mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}}, \boldsymbol{e}_{\mathrm{pa}(\tilde{\mathcal{A}})}) \\ \boldsymbol{x}_{\tilde{\mathcal{A}} \setminus \mathcal{C}} & = (\boldsymbol{g}_{\tilde{\mathcal{A}}})_{\tilde{\mathcal{A}} \setminus \mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}}, \boldsymbol{e}_{\mathrm{pa}(\tilde{\mathcal{A}})}) \end{cases} \\
\iff & \begin{cases} \boldsymbol{x}_{\mathcal{A}} & = \boldsymbol{g}_{\mathcal{A}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{A})}) \\ \boldsymbol{x}_{\tilde{\mathcal{A}}} & = \boldsymbol{g}_{\tilde{\mathcal{A}}}(\boldsymbol{x}_{\mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}}, \boldsymbol{e}_{\mathrm{pa}(\tilde{\mathcal{A}})}) . \end{cases}
\end{aligned}$$

Now $\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A} \subseteq \mathrm{pa}(\mathcal{D}) \setminus \mathcal{D}$, and similarly, $\mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}} \subseteq \mathrm{pa}(\mathcal{D}) \setminus \mathcal{D}$. Hence, we conclude that the mapping $\boldsymbol{h}_{\mathcal{D}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{D}) \setminus \mathcal{D}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{D})} \to \boldsymbol{\mathcal{X}}_{\mathcal{D}}$ defined by

$$\begin{aligned}
& \boldsymbol{h}_{\mathcal{D}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{D}) \setminus \mathcal{D}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{D})}) := \\
& \quad \big((\boldsymbol{g}_{\mathcal{A}})_{\mathcal{A} \setminus \mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{A})}), (\boldsymbol{g}_{\mathcal{A}})_{\mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{A})}), (\boldsymbol{g}_{\tilde{\mathcal{A}}})_{\tilde{\mathcal{A}} \setminus \mathcal{C}}(\boldsymbol{x}_{\mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}}, \boldsymbol{e}_{\mathrm{pa}(\tilde{\mathcal{A}})})\big)
\end{aligned}$$

is a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{D}$, and that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{D}$. □

PROOF OF COROLLARY 3.4.5. It suffices to show the implication to the left. We have to show that $\mathcal{M}$ is uniquely solvable w.r.t. each ancestral subset of $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$. The proof proceeds via induction with respect to the size of the ancestral subset. For ancestral subsets of size 0, the claim is trivially true. Ancestral subsets of size 1 must be of the form $\{i\} = \mathrm{an}_{\mathcal{G}(\mathcal{M})_\mathcal{O}}(i)$ for $i \in \mathcal{O}$ and hence the claim is true by assumption. Assume that the claim holds for all ancestral subsets of size $\leq n$. Let $\mathcal{A}$ be an ancestral subset of $\mathcal{G}(\mathcal{M})_\mathcal{O}$ of size $n+1$. If $\mathcal{A} = \mathrm{an}_{\mathcal{G}(\mathcal{M})_\mathcal{O}}(i)$ for some $i \in \mathcal{O}$ then the claim holds for $\mathcal{A}$ by assumption. Otherwise, $\mathcal{A} = \bigcup_{i\in\mathcal{A}} \mathrm{an}_{\mathcal{G}(\mathcal{M})_\mathcal{O}}(i)$ is a union of ancestral subsets of size $\leq n$. Choose distinct elements $\{i_1, \ldots, i_k\} \subseteq \mathcal{A}$ where k is the smallest integer such that $\bigcup_{j=1}^k \mathrm{an}_{\mathcal{G}(\mathcal{M})_\mathcal{O}}(i_j) = \mathcal{A}$. By applying Proposition 3.4.4 to $\bigcup_{j=1}^{k-1} \mathrm{an}_{\mathcal{G}(\mathcal{M})_\mathcal{O}}(i_j)$ and $\mathrm{an}_{\mathcal{G}(\mathcal{M})_\mathcal{O}}(i_k)$, thereby noting that the intersection of these two sets is an ancestral subset of size $\leq n$ and making use of the induction hypothesis, we arrive at the conclusion that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{A}$. $\square$

PROOF OF PROPOSITION 3.5.2. We first show that simplicity is preserved under perfect intervention. Let $\mathcal{M}$ be a simple SCM, $\mathcal{O} \subseteq \mathcal{I}$, $I \subseteq \mathcal{I}$ and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$. Define $\mathcal{O}_1 := \mathcal{O} \cap I$ and $\mathcal{O}_2 := \mathcal{O} \setminus I$, then $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2$. Note that $\mathrm{pa}(\mathcal{O}_2) \setminus \mathcal{O}_2 = (\mathrm{pa}(\mathcal{O}_2) \setminus (\mathcal{O}_2 \cup I)) \cup (\mathrm{pa}(\mathcal{O}_2) \cap I)$ and $\mathrm{pa}(\mathcal{O}_2) \setminus (\mathcal{O}_2 \cup I) \subseteq \mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}$. Let $\boldsymbol{g}_{\mathcal{O}_2} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}_2)\setminus\mathcal{O}_2} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O}_2)} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}_2}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{O}_2$. The mapping $\tilde{\boldsymbol{g}}_\mathcal{O} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_\mathcal{O}$ defined by

$$\begin{cases} (\tilde{\boldsymbol{g}}_\mathcal{O})_{\mathcal{O}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) := \boldsymbol{\xi}_{\mathcal{O}_1} \\ (\tilde{\boldsymbol{g}}_\mathcal{O})_{\mathcal{O}_2}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}) := \boldsymbol{g}_{\mathcal{O}_2}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O}_2)\setminus(\mathcal{O}_2\cup I)}, \boldsymbol{\xi}_{\mathrm{pa}(\mathcal{O}_2)\cap I}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O}_2)}) \end{cases}$$

is a measurable solution function for $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ w.r.t. $\mathcal{O}$, and it is clear that $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is uniquely solvable w.r.t. $\mathcal{O}$.

Next, we show that simplicity is preserved under the twin operation. Let $\tilde{\mathcal{O}} \subseteq \mathcal{I} \cup \mathcal{I}'$. Take $\mathcal{O}_1 = \tilde{\mathcal{O}} \cap \mathcal{I}$, $\mathcal{O}_2' = \tilde{\mathcal{O}} \cap \mathcal{I}'$ and $\mathcal{O}_2$ the original copy of $\mathcal{O}_2'$ in $\mathcal{I}$. Let $\boldsymbol{g}_{\mathcal{O}_1} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}_1)\setminus\mathcal{O}_1} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O}_1)} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}_1}$ and $\boldsymbol{g}_{\mathcal{O}_2} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}_2)\setminus\mathcal{O}_2} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O}_2)} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}_2}$ be measurable solution functions for $\mathcal{M}$ w.r.t. $\mathcal{O}_1$ and $\mathcal{O}_2$ respectively. Define now the mapping $\boldsymbol{h}_{\tilde{\mathcal{O}}} : \boldsymbol{\mathcal{X}}_{\widetilde{\mathrm{pa}}(\tilde{\mathcal{O}})\setminus\tilde{\mathcal{O}}} \times \boldsymbol{\mathcal{E}}_{\widetilde{\mathrm{pa}}(\tilde{\mathcal{O}})} \to \boldsymbol{\mathcal{X}}_{\tilde{\mathcal{O}}}$ by

$$\begin{aligned} (\boldsymbol{h}_{\tilde{\mathcal{O}}})_{\tilde{\mathcal{O}}\cap\mathcal{I}}(\boldsymbol{x}_{\widetilde{\mathrm{pa}}(\tilde{\mathcal{O}})\setminus\tilde{\mathcal{O}}}, \boldsymbol{e}_{\widetilde{\mathrm{pa}}(\tilde{\mathcal{O}})}) &:= \boldsymbol{g}_{\mathcal{O}_1}(\boldsymbol{x}_{\widetilde{\mathrm{pa}}(\mathcal{O}_1)\setminus\mathcal{O}_1}, \boldsymbol{e}_{\widetilde{\mathrm{pa}}(\mathcal{O}_1)}) \\ (\boldsymbol{h}_{\tilde{\mathcal{O}}})_{\tilde{\mathcal{O}}\cap\mathcal{I}'}(\boldsymbol{x}_{\widetilde{\mathrm{pa}}(\tilde{\mathcal{O}})\setminus\tilde{\mathcal{O}}}, \boldsymbol{e}_{\widetilde{\mathrm{pa}}(\tilde{\mathcal{O}})}) &:= \boldsymbol{g}_{\mathcal{O}_2}(\boldsymbol{x}_{\widetilde{\mathrm{pa}}(\mathcal{O}_2')\setminus\mathcal{O}_2'}, \boldsymbol{e}_{\widetilde{\mathrm{pa}}(\mathcal{O}_2')}), \end{aligned}$$

where we define $\widetilde{\mathrm{pa}} := \mathrm{pa}_{\mathcal{G}^a(\mathcal{M}^{\mathrm{twin}})}$ as the parents w.r.t. the twin graph $\mathcal{G}^a(\mathcal{M}^{\mathrm{twin}})$. Then by construction this mapping $\boldsymbol{h}_{\tilde{\mathcal{O}}}$ is a measurable solution function for $\mathcal{M}^{\mathrm{twin}}$ w.r.t. $\tilde{\mathcal{O}}$, and it is clear that $\mathcal{M}^{\mathrm{twin}}$ is uniquely solvable w.r.t. $\tilde{\mathcal{O}}$. $\square$

PROOF OF COROLLARY 3.5.3. From Proposition 3.5.2 it follows that the observational and all the intervened models of $\mathcal{M}$ and $\mathcal{M}^{\mathrm{twin}}$ are uniquely solvable. From Theorem 3.2.5 it now follows that $\mathcal{M}$ induces unique observational, interventional and counterfactual distributions. $\square$

PROOF OF PROPOSITION 4.3.2. The twin operation preserves the equivalence relation on SCMs and since equivalent SCMs are interventionally equivalent w.r.t. every subset, the two equivalent twin SCMs have to be interventionally equivalent w.r.t. $\mathcal{O} \cup \mathcal{O}'$ for every $\mathcal{O} \subseteq \mathcal{I}$ with $\mathcal{O}'$ the copy of $\mathcal{O}$ in $\mathcal{I}'$. $\square$

LEMMA C.2. *An SCM $\mathcal{M}$ is observationally equivalent to $\mathcal{M}^{\mathrm{twin}}$ w.r.t. $\mathcal{O} \subseteq \mathcal{I}$.*

PROOF. Let $(\boldsymbol{X}, \boldsymbol{E})$ be a solution of $\mathcal{M}$, then $((\boldsymbol{X}, \boldsymbol{X}), \boldsymbol{E})$ is a solution of $\mathcal{M}^{\mathrm{twin}}$. Conversely, let $((\boldsymbol{X}, \boldsymbol{X}'), \boldsymbol{E})$ be a solution of $\mathcal{M}^{\mathrm{twin}}$, then $(\boldsymbol{X}, \boldsymbol{E})$ is a solution of $\mathcal{M}$. $\square$

PROOF OF PROPOSITION 4.3.3. Let $\mathcal{M}$ and $\tilde{\mathcal{M}}$ be counterfactually equivalent w.r.t. $\mathcal{O}$. Then $\mathcal{M}^{\text{twin}}$ and $\tilde{\mathcal{M}}^{\text{twin}}$ are interventionally equivalent w.r.t. $\mathcal{O} \cup \mathcal{O}'$. Thus for $I \subseteq \mathcal{O}$, $I' \subseteq \mathcal{O}'$ the copy of I and $\boldsymbol{\xi}_{I'} = \boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, $\mathcal{M}^{\text{twin}}_{\text{do}(I \cup I', \boldsymbol{\xi}_{I \cup I'})}$ and $\tilde{\mathcal{M}}^{\text{twin}}_{\text{do}(I \cup I', \boldsymbol{\xi}_{I \cup I'})}$ are observationally equivalent w.r.t. $\mathcal{O} \cup \mathcal{O}'$. In particular, they are observationally equivalent w.r.t. $\mathcal{O}$. From Proposition 2.5.2 we have that $\mathcal{M}^{\text{twin}}_{\text{do}(I \cup I', \boldsymbol{\xi}_{I \cup I'})} = (\mathcal{M}_{\text{do}(I, \boldsymbol{\xi}_I)})^{\text{twin}}$ and $\tilde{\mathcal{M}}^{\text{twin}}_{\text{do}(I \cup I', \boldsymbol{\xi}_{I \cup I'})} = (\tilde{\mathcal{M}}_{\text{do}(I, \boldsymbol{\xi}_I)})^{\text{twin}}$, and together with Lemma C.2 this gives that $\mathcal{M}_{\text{do}(I, \boldsymbol{\xi}_I)}$ and $\tilde{\mathcal{M}}_{\text{do}(I, \boldsymbol{\xi}_I)}$ are observationally equivalent w.r.t. $\mathcal{O}$. □

The following lemma shows that it does not matter if a candidate measurable solution function for concluding unique solvability has "redundant" inputs.

LEMMA C.3. *Let $\mathcal{M}$ be an SCM. Let $B \subseteq \mathcal{I}$ and $A \subseteq \mathcal{I} \cup \mathcal{J}$ such that $(\text{pa}(B) \setminus B) \subseteq A$ and $B \cap A = \emptyset$. Assume that $\boldsymbol{g}_B : \boldsymbol{\mathcal{X}}_A \times \boldsymbol{\mathcal{E}}_A \to \boldsymbol{\mathcal{X}}_B$ is a measurable function such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*

$$\boldsymbol{x}_B = \boldsymbol{f}_B(\boldsymbol{x}_{\text{pa}(B)}, \boldsymbol{e}_{\text{pa}(B)}) \iff \boldsymbol{x}_B = \boldsymbol{g}_B(\boldsymbol{x}_A, \boldsymbol{e}_A)\,.$$

Then $\mathcal{M}$ is uniquely solvable w.r.t. B.

PROOF. Assume that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_B = \boldsymbol{f}_B(\boldsymbol{x}_{\text{pa}(B)}, \boldsymbol{e}_{\text{pa}(B)}) \iff \boldsymbol{x}_B = \boldsymbol{g}_B(\boldsymbol{x}_A, \boldsymbol{e}_A)\,.$$

Let $C := A \setminus (\text{pa}(B) \setminus B)$, then by Lemma D.11.(7) we have that there exists $\hat{\boldsymbol{e}}_C \in \boldsymbol{\mathcal{E}}_C$ and $\hat{\boldsymbol{x}}_C \in \boldsymbol{\mathcal{X}}_C$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathcal{J} \setminus C}}$-almost every $\boldsymbol{e}_{\mathcal{J} \setminus C} \in \boldsymbol{\mathcal{E}}_{\mathcal{J} \setminus C}$ and for all $\boldsymbol{x}_{\mathcal{I} \setminus C} \in \boldsymbol{\mathcal{X}}_{\mathcal{I} \setminus C}$

$$\boldsymbol{x}_B = \boldsymbol{f}_B(\boldsymbol{x}_{\text{pa}(B)}, \boldsymbol{e}_{\text{pa}(B)}) \iff \boldsymbol{x}_B = \boldsymbol{g}_B(\boldsymbol{x}_{\text{pa}(B) \setminus B}, \hat{\boldsymbol{x}}_C, \boldsymbol{e}_{\text{pa}(B)}, \hat{\boldsymbol{e}}_C)\,.$$

Defining the mapping $\boldsymbol{h}_B : \boldsymbol{\mathcal{X}}_{\text{pa}(B) \setminus B} \times \boldsymbol{\mathcal{E}}_{\text{pa}(B)} \to \boldsymbol{\mathcal{X}}_B$ by

$$\boldsymbol{h}_B(\boldsymbol{x}_{\text{pa}(B) \setminus B}, \boldsymbol{e}_{\text{pa}(B)}) := \boldsymbol{g}_B(\boldsymbol{x}_{\text{pa}(B) \setminus B}, \hat{\boldsymbol{x}}_C, \boldsymbol{e}_{\text{pa}(B)}, \hat{\boldsymbol{e}}_C)\,,$$

where we picked $\hat{\boldsymbol{e}}_C \in \boldsymbol{\mathcal{E}}_C$ and $\hat{\boldsymbol{x}}_C \in \boldsymbol{\mathcal{X}}_C$ such that the above equivalence holds, and applying Lemma D.11.(6) we get that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_B = \boldsymbol{f}_B(\boldsymbol{x}_{\text{pa}(B)}, \boldsymbol{e}_{\text{pa}(B)}) \iff \boldsymbol{x}_B = \boldsymbol{h}_B(\boldsymbol{x}_{\text{pa}(B) \setminus B}, \boldsymbol{e}_{\text{pa}(B)})$$

holds. Thus, $\mathcal{M}$ is uniquely solvable w.r.t. B. □

PROOF OF PROPOSITION 5.1.4. From unique solvability of $\mathcal{M}$ w.r.t. $\mathcal{L}_1$ it follows that there exists a mapping $\boldsymbol{g}_{\mathcal{L}_1} : \boldsymbol{\mathcal{X}}_{\text{pa}(\mathcal{L}_1) \setminus (\mathcal{L}_1)} \times \boldsymbol{\mathcal{E}}_{\text{pa}(\mathcal{L}_1)} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_1}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\text{pa}(\mathcal{L}_1) \setminus \mathcal{L}_1}, \boldsymbol{e}_{\text{pa}(\mathcal{L}_1)}) \quad \iff \quad \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{f}_{\mathcal{L}_1}(\boldsymbol{x}, \boldsymbol{e})\,.$$

Let $\widehat{\text{pa}}$ denotes the parents in $\mathcal{G}^a(\mathcal{M}_{\text{marg}(\mathcal{L}_1)})$. Note that $\widehat{\text{pa}}(\mathcal{L}_2) \setminus \mathcal{L}_2 \subseteq \text{pa}(\mathcal{L}_1 \cup \mathcal{L}_2) \setminus (\mathcal{L}_1 \cup \mathcal{L}_2)$. Let $\tilde{\boldsymbol{f}}$ denote the marginal causal mechanism of a structurally minimal SCM that is equivalent to the marginalization $\mathcal{M}_{\text{marg}(\mathcal{L}_1)}$ constructed from $\boldsymbol{g}_{\mathcal{L}_1}$ (see Proposition 2.3.3).

$\Longrightarrow$: If $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$, then there exists a mapping $\tilde{\boldsymbol{g}}_{\mathcal{L}_2} : \boldsymbol{\mathcal{X}}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2} \times \boldsymbol{\mathcal{E}}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_2}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\mathcal{I}\setminus\mathcal{L}_1} \in \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\mathcal{L}_1}$

$$\boldsymbol{x}_{\mathcal{L}_2} = \tilde{\boldsymbol{g}}_{\mathcal{L}_2}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}) \iff \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}), \boldsymbol{x}_{\mathcal{I}\setminus\mathcal{L}_1}, \boldsymbol{e})\,.$$

Define the mapping $\boldsymbol{h} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_1\cup\mathcal{L}_2}$ by

$$\begin{aligned}&(\boldsymbol{h}_{\mathcal{L}_1}, \boldsymbol{h}_{\mathcal{L}_2})(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)}) := \\ &\Big(\boldsymbol{g}_{\mathcal{L}_1}\big((\tilde{\boldsymbol{g}}_{\mathcal{L}_2})_{\mathrm{pa}(\mathcal{L}_1)}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}), \boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}\big), \tilde{\boldsymbol{g}}_{\mathcal{L}_2}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)})\Big)\,.\end{aligned}$$

Then for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\begin{aligned}
&\begin{cases}\boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{f}_{\mathcal{L}_1}(\boldsymbol{x}, \boldsymbol{e})\\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{x}, \boldsymbol{e})\end{cases}\\
\iff &\begin{cases}\boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)})\\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{x}, \boldsymbol{e})\end{cases}\\
\iff &\begin{cases}\boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)})\\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}), \boldsymbol{x}_{\mathcal{I}\setminus\mathcal{L}_1}, \boldsymbol{e})\end{cases}\\
\iff &\begin{cases}\boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)})\\ \boldsymbol{x}_{\mathcal{L}_2} &= \tilde{\boldsymbol{g}}_{\mathcal{L}_2}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)})\end{cases}\\
\iff &\begin{cases}\boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}\big((\tilde{\boldsymbol{g}}_{\mathcal{L}_2})_{\mathrm{pa}(\mathcal{L}_1)}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}), \boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}\big)\\ \boldsymbol{x}_{\mathcal{L}_2} &= \tilde{\boldsymbol{g}}_{\mathcal{L}_2}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)\setminus\mathcal{L}_2}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)})\end{cases}\\
\iff &\begin{cases}\boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{h}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)})\\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{h}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)})\,,\end{cases}
\end{aligned}$$

where in the first equivalence we used unique solvability w.r.t. $\mathcal{L}_1$ of $\mathcal{M}$, in the second we used substitution, in the third we used unique solvability w.r.t. $\mathcal{L}_2$ of $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$, in the fourth we used again substitution and in the last equivalence we used the definition of $\boldsymbol{h}$. From this we conclude that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1 \cup \mathcal{L}_2$. Hence, by definition it follows that $\mathrm{marg}(\mathcal{L}_2)\circ\mathrm{marg}(\mathcal{L}_1)(\mathcal{M}) \equiv \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{M})$.

$\Longleftarrow$: If $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1 \cup \mathcal{L}_2$, there exists a mapping $\boldsymbol{h} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)} \times \boldsymbol{\mathcal{E}}_{\mathcal{L}_1\cup\mathcal{L}_2} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_1\cup\mathcal{L}_2}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{L}_1\cup\mathcal{L}_2} = \boldsymbol{h}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)}) \quad\iff\quad \boldsymbol{x}_{\mathcal{L}_1\cup\mathcal{L}_2} = \boldsymbol{f}_{\mathcal{L}_1\cup\mathcal{L}_2}(\boldsymbol{x}, \boldsymbol{e})\,.$$

Then, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$
\begin{aligned}
&\begin{cases} \boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{h}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)}) \\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{h}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)}) \end{cases} \\
\iff &\begin{cases} \boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{f}_{\mathcal{L}_1}(\boldsymbol{x}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{x}, \boldsymbol{e}) \end{cases} \\
\iff &\begin{cases} \boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}) \\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{x}, \boldsymbol{e}) \end{cases} \\
\iff &\begin{cases} \boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}) \\ \boldsymbol{x}_{\mathcal{L}_2} &= \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}), \boldsymbol{x}_{\mathcal{I}\setminus\mathcal{L}_1}, \boldsymbol{e}) \end{cases} \\
\iff &\begin{cases} \boldsymbol{x}_{\mathcal{L}_1} &= \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1)\setminus\mathcal{L}_1}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1)}) \\ \boldsymbol{x}_{\mathcal{L}_2} &= \tilde{\boldsymbol{f}}_{\mathcal{L}_2}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}) \,. \end{cases}
\end{aligned}
$$

This gives for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ for all $\boldsymbol{x}_{\mathcal{I}\setminus\mathcal{L}_1} \in \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\mathcal{L}_1}$

$$
\begin{aligned}
&\boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{h}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)\setminus(\mathcal{L}_1\cup\mathcal{L}_2)}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L}_1\cup\mathcal{L}_2)}) \\
\iff &\boldsymbol{x}_{\mathcal{L}_2} = \tilde{\boldsymbol{f}}_{\mathcal{L}_2}(\boldsymbol{x}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}, \boldsymbol{e}_{\widehat{\mathrm{pa}}(\mathcal{L}_2)}) \,.
\end{aligned}
$$

Now apply Lemma C.3 to conclude that $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$. □

Proof of Proposition 5.1.5. This follows straightforwardly from the definitions of perfect intervention and marginalization and the fact that if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is also uniquely solvable w.r.t. $\mathcal{L}$, since the structural equations for the variables $\mathcal{L}$ are the same for $\mathcal{M}$ and $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$. □

Proof of Proposition 5.1.6. This follows straightforwardly from the definition of the twin operation and marginalization and the fact that if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then $\mathrm{twin}(\mathcal{M})$ is uniquely solvable w.r.t. $\mathcal{L} \cup \mathcal{L}'$, where $\mathcal{L}'$ is the copy of $\mathcal{L}$ in $\mathcal{I}'$. □

Lemma C.4. *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L} \subsetneq \mathcal{I}$ such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$. Then $\mathcal{M}$ and $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ are observationally equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.*

Proof. Let $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$. From unique solvability w.r.t. $\mathcal{L}$ it follows that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$
\begin{aligned}
&\begin{cases} \boldsymbol{x}_{\mathcal{L}} &= \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{O}} &= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e}) \end{cases} \\
\iff &\begin{cases} \boldsymbol{x}_{\mathcal{L}} &= \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}) \\ \boldsymbol{x}_{\mathcal{O}} &= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}), \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \\
\iff &\begin{cases} \boldsymbol{x}_{\mathcal{L}} &= \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}) \\ \boldsymbol{x}_{\mathcal{O}} &= \tilde{\boldsymbol{f}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \,, \end{cases}
\end{aligned}
$$

where $\tilde{\boldsymbol{f}}$ is the marginal causal mechanism of $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ constructed from a measurable solution function $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ for $\mathcal{M}$ w.r.t. $\mathcal{L}$. Hence, a solution $(\boldsymbol{X}, \boldsymbol{E})$ of $\mathcal{M}$ satisfies $\boldsymbol{X}_{\mathcal{O}} = \tilde{\boldsymbol{f}}(\boldsymbol{X}_{\mathcal{O}}, \boldsymbol{E})$ a.s.. Conversely, if $(\tilde{\boldsymbol{X}}_{\mathcal{O}}, \boldsymbol{E})$ is a solution of the marginal SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ then with $\tilde{\boldsymbol{X}}_{\mathcal{L}} := \boldsymbol{g}_{\mathcal{L}}(\tilde{\boldsymbol{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}}, \boldsymbol{E}_{\mathrm{pa}(\mathcal{L})})$, the random variables $(\boldsymbol{X}, \boldsymbol{E}) := (\tilde{\boldsymbol{X}}_{\mathcal{O}}, \tilde{\boldsymbol{X}}_{\mathcal{L}}, \boldsymbol{E})$ are a solution of $\mathcal{M}$. □

Proof of Theorem 5.1.7. The observational equivalence follows from Lemma C.4. Using both Lemma C.4 and Proposition 5.1.5 we can prove the interventional equivalence. Observe that from Proposition 5.1.5 we know that for a subset $I \subseteq \mathcal{I}\setminus\mathcal{L}$ and a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, $(\mathrm{marg}(\mathcal{L})\circ\mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M})$ exists. By Lemma C.4 we know that $\mathrm{do}(I, \boldsymbol{\xi}_I)(\mathcal{M})$ and $(\mathrm{marg}(\mathcal{L})\circ\mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M})$ are observationally equivalent w.r.t. $\mathcal{O}$ and hence by applying again Proposition 5.1.5, $\mathrm{do}(I, \boldsymbol{\xi}_I)(\mathcal{M})$ and $(\mathrm{do}(I, \boldsymbol{\xi}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$ are observationally equivalent w.r.t. $\mathcal{O}$. This implies that $\mathcal{M}$ and $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ are interventionally equivalent w.r.t. $\mathcal{O}$. Lastly, we need to show that $\mathrm{twin}(\mathcal{M})$ and $(\mathrm{twin}\circ\mathrm{marg}(\mathcal{L}))(\mathcal{M})$ are interventionally equivalent w.r.t. $(\mathcal{I} \cup \mathcal{I}') \setminus (\mathcal{L} \cup \mathcal{L}')$, where $\mathcal{L}'$ is the copy of $\mathcal{L}$ in $\mathcal{I}'$. From Proposition 5.1.6 $(\mathrm{twin}\circ\mathrm{marg}(\mathcal{L}))(\mathcal{M})$ is equivalent to $(\mathrm{marg}(\mathcal{L}\cup\mathcal{L}')\circ\mathrm{twin})(\mathcal{M})$ and since we proved that $(\mathrm{marg}(\mathcal{L}\cup\mathcal{L}') \circ \mathrm{twin})(\mathcal{M})$ and $\mathrm{twin}(\mathcal{M})$ are interventionally equivalent w.r.t. $(\mathcal{I}\cup\mathcal{I}') \setminus (\mathcal{L}\cup\mathcal{L}')$ the result follows. □

Proof of Proposition 5.2.2. A similar proof as for Theorem 1 in (Evans, 2016) works. □

Proof of Proposition 5.2.3. Applying the $\mathrm{do}(I)$ operation to the latent projection $\mathrm{marg}(\mathcal{L})(\mathcal{G})$ removes all the incoming edges on the nodes I. Such an incoming edge at a node in I in $\mathrm{marg}(\mathcal{L})(\mathcal{G})$ corresponds to a path in $\mathcal{G}$ that points to that node. But since $\mathrm{do}(I)(\mathcal{G})$ is just $\mathcal{G}$ with all the incoming edges on I removed, the graph $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I))(\mathcal{G})$ also has all the incoming edges on the nodes I removed. □

Proof of Proposition 5.2.4. We will denote the copy in I' of any node $i \in I$ by i', i.e., $I' = \{i' : i \in I\}$. The edges in $(\mathrm{twin}(I \setminus \mathcal{L}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})$ can be partitioned into three cases:

$$\begin{cases} v \to w & v \in J \cup I \setminus \mathcal{L}, w \in J \cup I \setminus \mathcal{L}, v \to w \in \mathrm{marg}(\mathcal{L})(\mathcal{G}) \\ v \to w' & v \in J, w \in I \setminus \mathcal{L}, v \to w \in \mathrm{marg}(\mathcal{L})(\mathcal{G}), \\ v' \to w' & v \in I \setminus \mathcal{L}, w \in I \setminus \mathcal{L}, v \to w \in \mathrm{marg}(\mathcal{L})(\mathcal{G}), \end{cases}$$

where $J := \mathcal{V} \setminus I$.

Note that in $\mathrm{twin}(I)(\mathcal{G})$, there are no directed edges of the form $v' \to w$ by definition. Therefore, the edges in $(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin}(I))(\mathcal{G})$ can be partitioned into three cases:

$$\begin{cases} v \to w & v \in J \cup I \setminus \mathcal{L}, w \in J \cup I \setminus \mathcal{L}, v \to \ell_1 \to \cdots \to \ell_n \to w \in \mathrm{twin}(I)(\mathcal{G}), \\ v \to w' & v \in J, w \in I \setminus \mathcal{L}, v \to \ell'_1 \to \cdots \to \ell'_n \to w' \in \mathrm{twin}(I)(\mathcal{G}) \\ v' \to w' & v \in I \setminus \mathcal{L}, w \in I \setminus \mathcal{L}, v' \to \ell'_1 \to \cdots \to \ell'_n \to w' \in \mathrm{twin}(I)(\mathcal{G}), \end{cases}$$

where all $\ell_1, \ldots, \ell_n \in \mathcal{L}$ and $\ell'_1, \ldots, \ell'_n \in \mathcal{L}'$. Thus, the non-endpoint nodes on the directed paths in $\mathrm{twin}(I)(\mathcal{G})$ must either all lie in $\mathcal{L}$ or in $\mathcal{L}'$. With the definition of $\mathrm{twin}(I)(\mathcal{G})$ we can rewrite this as follows:

$$\begin{cases} v \to w & v \in J \cup I \setminus \mathcal{L}, w \in J \cup I \setminus \mathcal{L}, v \to \ell_1 \to \cdots \to \ell_n \to w \in \mathcal{G}, \\ v \to w' & v \in J, w \in I \setminus \mathcal{L}, v \to \ell_1 \to \cdots \to \ell_n \to w \in \mathcal{G} \\ v' \to w' & v \in I \setminus \mathcal{L}, w \in I \setminus \mathcal{L}, v \to \ell_1 \to \cdots \to \ell_n \to w \in \mathcal{G}, \end{cases}$$

where all intermediate $\ell_1, \ldots, \ell_n$ must lie in $\mathcal{L}$. This corresponds exactly with the edges in $(\mathrm{twin}(I \setminus \mathcal{L}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})$. □

Proof of Proposition 5.2.6. Without loss of generality, we assume that $\mathcal{M}$ is structurally minimal (see Proposition 2.3.3). Let $\boldsymbol{g}_{\mathcal{L}}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{L}$ and denote by $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ the marginal SCM constructed from $\boldsymbol{g}_{\mathcal{L}}$. For $j \in \mathcal{I} \setminus \mathcal{L}$, define $A_j := \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{L}}}(\mathrm{pa}(j) \cap \mathcal{L}) \subseteq \mathcal{L}$ and let $\tilde{\boldsymbol{g}}_{A_j}$ be a measurable solution function for $\mathcal{M}$ w.r.t. A_j. Because $A_j \subseteq \mathcal{L}$ and $\mathrm{pa}(A_j) \setminus A_j \subseteq \mathrm{pa}(\mathcal{L}) \setminus \mathcal{L}$, by Lemma C.1, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$(\boldsymbol{g}_{\mathcal{L}})_{A_j}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}) \setminus \mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}) = \tilde{\boldsymbol{g}}_{A_j}(\boldsymbol{x}_{\mathrm{pa}(A_j) \setminus A_j}, \boldsymbol{e}_{\mathrm{pa}(A_j)}) .$$

Therefore, the component $\tilde{f}_j$ of the marginal causal mechanism $\tilde{\boldsymbol{f}}$ of $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ satisfies for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\begin{aligned} \tilde{f}_j(\boldsymbol{x}_{\mathcal{I} \setminus \mathcal{L}}, \boldsymbol{e}) &:= f_j\big((\boldsymbol{g}_{\mathcal{L}})_{\mathrm{pa}(j)}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{L}) \setminus \mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{L})}), \boldsymbol{x}_{\mathrm{pa}(j) \setminus \mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(j)}\big) \\ &= f_j\big((\tilde{\boldsymbol{g}}_{A_j})_{\mathrm{pa}(j) \cap \mathcal{L}}(\boldsymbol{x}_{\mathrm{pa}(A_j) \setminus A_j}, \boldsymbol{e}_{\mathrm{pa}(A_j)}), \boldsymbol{x}_{\mathrm{pa}(j) \setminus \mathcal{L}}, \boldsymbol{e}_{\mathrm{pa}(j)}\big) . \end{aligned}$$

Hence, the endogenous parents of j in $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ are a subset of $\big((\mathrm{pa}(A_j) \setminus A_j) \cup (\mathrm{pa}(j) \setminus \mathcal{L})\big) \cap \mathcal{I}$ and the exogenous parents of j in $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ are a subset of $(\mathrm{pa}(A_j) \cup \mathrm{pa}(j)) \cap \mathcal{J}$. Hence, all parents of j in $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ are a subset of those $k \in (\mathcal{I} \setminus \mathcal{L}) \cup \mathcal{J}$ such that there exists a path $k \to \ell_1 \to \cdots \to \ell_n \to j \in \mathcal{G}^a(\mathcal{M})$ for $n \geq 0$ and $\ell_1, \ldots, \ell_n \in \mathcal{L}$. Therefore, the augmented graph $\mathcal{G}^a\big(\mathrm{marg}(\mathcal{L})(\mathcal{M})\big)$ is a subgraph of the latent projection $\mathrm{marg}(\mathcal{L})\big(\mathcal{G}^a(\mathcal{M})\big)$. Hence,

$$\begin{aligned} \mathcal{G}\big(\mathrm{marg}(\mathcal{L})(\mathcal{M})\big) &= \mathrm{marg}(\mathcal{J})\Big(\mathcal{G}^a\big(\mathrm{marg}(\mathcal{L})(\mathcal{M})\big)\Big) \\ &\subseteq \mathrm{marg}(\mathcal{J})\Big(\mathrm{marg}(\mathcal{L})\big(\mathcal{G}^a(\mathcal{M})\big)\Big) \\ &= \mathrm{marg}(\mathcal{L})\Big(\mathrm{marg}(\mathcal{J})\big(\mathcal{G}^a(\mathcal{M})\big)\Big) \\ &= \mathrm{marg}(\mathcal{L})\big(\mathcal{G}(\mathcal{M})\big) \end{aligned}$$

and we conclude that also the graph $\mathcal{G}\big(\mathrm{marg}(\mathcal{L})(\mathcal{M})\big)$ is a subgraph of the latent projection $\mathrm{marg}(\mathcal{L})\big(\mathcal{G}(\mathcal{M})\big)$. □

Proof of Proposition 5.2.8. Take two disjoint subsets $\mathcal{L}_1$ and $\mathcal{L}_2$ in $\mathcal{I}$. Then, it suffices to show that $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$. This follows directly from Proposition 5.1.4. □

Proof of Theorem 6.1.1. The first case is a well-known result. An elementary proof is obtained by noting that an acyclic system of structural equations trivially satisfies the local directed Markov property, and then apply (Lauritzen et al., 1990, Proposition 4), followed by applying the stability of d-separation with respect to (graphical) marginalization (Forré and Mooij, 2017, Lemma

2.2.15). Alternatively, the result also follows from sequential application of Theorems 3.8.2, 3.8.11, 3.7.7, 3.7.2 and 3.3.3 (using Remark 3.3.4) in (Forré and Mooij, 2017).

The discrete case is proved by the series of results Theorem 3.8.12, Remark 3.7.2, Theorem 3.6.6 and 3.5.2 in (Forré and Mooij, 2017).

The linear case is proved in Example 3.8.17 in (Forré and Mooij, 2017). To connect the assumptions made there with the ones we state here, observe that under the linear transformation rule for Lebesgue measures, the image measure of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ under the linear mapping $\mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{I}} : \boldsymbol{e} \mapsto \Gamma_{\mathcal{I}\mathcal{J}}\boldsymbol{e}$ gives a measure on $\boldsymbol{\mathcal{X}} = \mathbb{R}^{\mathcal{I}}$ with a density w.r.t. the Lebesgue measure on $\mathbb{R}^{\mathcal{I}}$, as long as the image of the linear mapping is the entire $\mathbb{R}^{\mathcal{I}}$. This is guaranteed if each causal mechanism has a non-trivial dependence on some exogenous variable(s), i.e., for each $i \in \mathcal{I}$ there is some $j \in \mathcal{J}$ with $\Gamma_{ij} \neq 0$. □

Proof of Proposition 6.2.3. This follows directly from the fact that the strongly connected components of $\mathcal{G}^a(\mathcal{M})$ form a DAG by Lemma A.1.2 and that the directed edges in $\mathcal{G}^a(\mathrm{acy}(\mathcal{M}))$ by construction respect every topological ordering of that DAG. Both SCMs are observationally equivalent by construction. □

Proof of Proposition 6.2.5. This follows immediately from the Definitions 6.2.2 and 6.2.4. □

Proof of Lemma 6.2.8. It suffices to show that for every C-σ-open walk between i and j in $\mathcal{G}$, there exists a C-σ-open path between i and j in $\mathcal{G}$. Let $\pi = (i = i_0, \dots, i_n = j)$ be a C-σ-open walk in $\mathcal{G}$. If a node ℓ occurs more than once in π, let i_j be the first node in π and i_k the last node in π that are in the same strongly connected component as ℓ. Since i_j and i_k are in the same strongly connected component, there are directed paths $i_j \to \cdots \to i_k$ and $i_k \to \cdots \to i_j$ in $\mathcal{G}$. We now construct a new walk π' from π by replacing the subwalk between i_j and i_k of π by a particular directed path between i_j and i_k: (i) If $k = n$, or if $k < n$ and $i_k \to i_{k+1}$ on π, we replace it by a shortest directed path $i_j \to \cdots \to i_k$, otherwise (ii) we replace it by a shortest directed path $i_j \leftarrow \cdots \leftarrow i_k$. We now show that the new walk π' is still C-σ-open.

π' cannot become C-σ-blocked through one of the initial nodes $i_0 \dots i_{j-1}$ or one of the final nodes $i_{k+1} \dots i_n$ on π', since these nodes occur in the same local configuration on π and do not C-σ-block π by assumption. Furthermore, π' cannot become C-σ-blocked through one of the nodes strictly between i_j and i_k on π' (if there are any), since these nodes are all non-endpoint non-colliders that only point to nodes in the same strongly connected component on π'. Because π is C-σ-open, $i_k \notin C$ if $k = n$ or if $i_k \to i_{k+1}$ on π. This holds in particular in case (i). Similarly, $i_j \notin C$ if $j = 0$ or $i_{j-1} \leftarrow i_j$ on π.

In case (i), π' is not C-σ-blocked by i_k because i_k is a non-collider on π' but $i_k \notin C$. Also i_j does not C-σ-block π'. Assume $i_j \neq i_k$ (otherwise there is nothing to prove). If $j = 0$, or if $j > 0$ and $i_{j-1} \leftarrow i_j$ on π', then the same holds for π and hence $i_j \notin C$; i_j is then a non-collider on π', but $i_j \notin C$. If $j > 0$ and $i_{j-1} \leftrightarrow i_j$ or $i_{j-1} \to i_j$ on π' then i_j is a non-endpoint non-collider on π' that does not point to a node in another strongly connected component.

Now consider case (ii). If $j = 0$ or $i_{j-1} \leftarrow i_j$ on π' then this case is analogous to case (i). So assume $j > 0$ and $i_{j-1} \to i_j$ or $i_{j-1} \leftrightarrow i_j$ on π'. If i_j is an endpoint of π', then $i_j = i_k$ and $k = n$ and therefore $i_k \notin C$, and hence i_j and i_k do not C-σ-block π'. Otherwise, i_j must be a collider on π' (whether $i_j = i_k$ or not). Then on the subwalk of π between i_j and i_k there must be a directed path from i_j to a collider that is ancestor of C, which implies that i_j is itself ancestor of C, and hence i_j does not C-σ-block π'. Also i_k cannot C-σ-block π'. Assume $i_j \neq i_k$ (otherwise there is nothing to prove). Since $i_k \leftarrow i_{k+1}$ or $i_k \leftrightarrow i_{k+1}$ on π', i_k is a non-endpoint non-collider on π' that does not point to a node in another strongly connected component.

Now in π', the number of nodes that occurs more than once is at least one less than in π. Repeat this procedure until no nodes occur more than once. □

Proof of Proposition 6.2.10. This follows directly as a special case of Corollary 2.8.4 in (Forré and Mooij, 2017). □

Proof of Theorem 6.2.12. An SCM $\mathcal{M}$ that is uniquely solvable w.r.t. each strongly connected component is uniquely solvable and hence, by Theorem 3.2.5, all its solutions have the same observational distribution. The last statement follows from the series of results Theorem 3.8.2, 3.8.11, Lemma 3.7.7 and Remark 3.7.2 in (Forré and Mooij, 2017). Alternatively, we give here a shorter proof: Under the stated conditions one can always construct the acyclification $\mathrm{acy}(\mathcal{M})$ which is observationally equivalent to $\mathcal{M}$ and is acyclic (see Proposition 6.2.3) and hence we can apply Theorem 6.1.1 to $\mathrm{acy}(\mathcal{M})$. Together with Proposition 6.2.5 and 6.2.10 this gives

$$A \underset{\mathcal{G}(\mathcal{M})}{\overset{\sigma}{\perp}} B \,|\, C \iff A \underset{\mathrm{acy}(\mathcal{G}(\mathcal{M}))}{\overset{d}{\perp}} B \,|\, C \implies A \underset{\mathcal{G}(\mathrm{acy}(\mathcal{M}))}{\overset{d}{\perp}} B \,|\, C \implies \boldsymbol{X}_A \underset{\mathbb{P}_{\mathcal{M}}^{\boldsymbol{X}}}{\perp\!\!\!\perp} \boldsymbol{X}_B \,|\, \boldsymbol{X}_C \,,$$

for $A, B, C \subseteq \mathcal{I}$ and $\boldsymbol{X}$ a solution of $\mathcal{M}$. □

Proof of Corollary 6.2.13. First observe that simplicity is preserved under both intervention and the twin-operation (see Proposition 3.5.2). Now the first statement follows from Theorem 6.2.12 if one takes into account the identities of Proposition 2.4.3 and 2.5.4. Similarly, the last statement follows from Theorem 6.1.1. □

Proof of Proposition 7.1.1. Define $\tilde{\mathcal{M}} := \mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$, $\widetilde{\mathrm{pa}} := \mathrm{pa}_{\mathcal{G}^a(\tilde{\mathcal{M}})}$ and let $\mathcal{A} := \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})_{\setminus i}}(j)$. Suppose that $i \to j \notin \mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$ and assume that the two induced distributions do not coincide. Because $i \to j \notin \mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$ it follows that $(\widetilde{\mathrm{pa}}(\mathcal{A}) \setminus \mathcal{A}) \cap \mathcal{I} = \emptyset$. Let now $\tilde{\boldsymbol{g}}_{\mathcal{A}} : \boldsymbol{\mathcal{E}}_{\widetilde{\mathrm{pa}}(\mathcal{A})} \to \boldsymbol{\mathcal{X}}_{\mathcal{A}}$ be a measurable solution function for $\tilde{\mathcal{M}}$ w.r.t. $\mathcal{A}$, i.e., we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{A}} = \tilde{\boldsymbol{f}}_{\mathcal{A}}(\boldsymbol{x}, \boldsymbol{e}) \quad \iff \quad \boldsymbol{x}_{\mathcal{A}} = \tilde{\boldsymbol{g}}_{\mathcal{A}}(\boldsymbol{e}_{\widetilde{\mathrm{pa}}(\mathcal{A})}) \,,$$

where $\tilde{\boldsymbol{f}}$ is the marginal causal mechanism of $\tilde{\mathcal{M}}$. Because $i \notin \mathcal{A}$ and $j \in \mathcal{A}$, it follows that for the intervened model $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}$ the marginal solution X_j is also a marginal solution of $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\tilde{\xi}_i)}$ and vice versa, which is in contradiction with the assumption. □

Proof of Proposition 7.2.2. Define $\tilde{\mathcal{M}} := \mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$, $\widetilde{\mathrm{pa}} := \mathrm{pa}_{\mathcal{G}^a(\tilde{\mathcal{M}})}$, $\mathcal{A}_i := \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})}(i)$ and $\mathcal{A}_j^{\setminus i} := \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})_{\setminus i}}(j)$. Suppose that there does not exist a bidirected edge $i \leftrightarrow j$ in the latent projection $\mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$. Because $i \leftrightarrow j \notin \mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\tilde{\mathcal{M}}))$, where here $\tilde{\mathcal{M}}$ is the intervened model $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$, we have that $\mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus j}}(i) \cap \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus i}}(j) \cap \mathcal{J} = \emptyset$. From $j \notin \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})}(i)$ it follows that $\mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})_{\setminus j}}(i) = \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})}(i)$, and hence $\mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})}(i) \cap \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus i}}(j) \cap \mathcal{J} = \emptyset$. Observe that $\widetilde{\mathrm{pa}}(\mathcal{A}_i) \subseteq \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})}(i)$ and $\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \subseteq \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus i}}(j) \cup \{i\}$, and thus $\widetilde{\mathrm{pa}}(\mathcal{A}_i) \cap \widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \cap \mathcal{J} = \emptyset$. Let

$\boldsymbol{g}_{\mathcal{A}_i} : \boldsymbol{\mathcal{E}}_{\widetilde{\mathrm{pa}}(\mathcal{A}_i)} \to \boldsymbol{\mathcal{X}}_{\mathcal{A}_i}$ be a measurable solution function for $\tilde{\mathcal{M}}$ w.r.t. $\mathcal{A}_i$, i.e., we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{A}_i} = \tilde{\boldsymbol{f}}_{\mathcal{A}_i}(\boldsymbol{x}, \boldsymbol{e}) \quad \Longleftrightarrow \quad \boldsymbol{x}_{\mathcal{A}_i} = \boldsymbol{g}_{\mathcal{A}_i}(\boldsymbol{e}_{\widetilde{\mathrm{pa}}(\mathcal{A}_i)}) ,$$

where $\tilde{\boldsymbol{f}}$ is the intervened causal mechanism of $\tilde{\mathcal{M}}$. Because $\widetilde{\mathrm{pa}}(\mathcal{A}_i) \cap \widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \cap \mathcal{J} = \emptyset$ and $i \in \mathcal{A}_i$, we have that $X_i \perp\!\!\!\perp \boldsymbol{E}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}$ for every solution $(\boldsymbol{X}, \boldsymbol{E})$ of $\tilde{\mathcal{M}}$.

Assume for the moment that $i \in \widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \setminus \mathcal{A}_j^{\setminus i}$, then $(\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \setminus \mathcal{A}_j^{\setminus i}) \cap \mathcal{I} = \{i\}$. Let $\boldsymbol{g}_{\mathcal{A}_j^{\setminus i}} : \mathcal{X}_i \times \boldsymbol{\mathcal{E}}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})} \to \boldsymbol{\mathcal{X}}_{\mathcal{A}_j^{\setminus i}}$ be a measurable solution function for $\tilde{\mathcal{M}}$ w.r.t. $\mathcal{A}_j^{\setminus i}$, i.e., we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\boldsymbol{x}_{\mathcal{A}_j^{\setminus i}} = \tilde{\boldsymbol{f}}_{\mathcal{A}_j^{\setminus i}}(\boldsymbol{x}, \boldsymbol{e}) \iff \boldsymbol{x}_{\mathcal{A}_j^{\setminus i}} = \boldsymbol{g}_{\mathcal{A}_j^{\setminus i}}(x_i, \boldsymbol{e}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}) .$$

For every measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ there exists a version of the regular conditional probability $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B} \,|\, X_i = \xi_i)$ such that for every value $\xi_i \in \mathcal{X}_i$ it satisfies

$$\begin{aligned}
\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}\big(X_j \in \mathcal{B}_j \,|\, X_i = \xi_i\big) &= \mathbb{P}_{\tilde{\mathcal{M}}}\big(X_j \in \mathcal{B}_j \,|\, X_i = \xi_i\big) \\
&= \mathbb{P}_{\tilde{\mathcal{M}}}\big((\boldsymbol{g}_{\mathcal{A}_j^{\setminus i}})_j(X_i, \boldsymbol{E}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}) \in \mathcal{B}_j \,|\, X_i = \xi_i\big) \\
&= \mathbb{P}_{\tilde{\mathcal{M}}}\big((\boldsymbol{g}_{\mathcal{A}_j^{\setminus i}})_j(\xi_i, \boldsymbol{E}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}) \in \mathcal{B}_j \,|\, X_i = \xi_i\big) \\
&= \mathbb{P}_{\tilde{\mathcal{M}}}\big((\boldsymbol{g}_{\mathcal{A}_j^{\setminus i}})_j(\xi_i, \boldsymbol{E}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}) \in \mathcal{B}_j\big) \\
&= \mathbb{P}_{\tilde{\mathcal{M}}_{\mathrm{do}(\{i\},\xi_i)}}\big((\boldsymbol{g}_{\mathcal{A}_j^{\setminus i}})_j(X_i, \boldsymbol{E}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}) \in \mathcal{B}_j\big) \\
&= \mathbb{P}_{\tilde{\mathcal{M}}_{\mathrm{do}(\{i\},\xi_i)}}\big(X_j \in \mathcal{B}_j\big) \\
&= \mathbb{P}_{\left(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}\right)_{\mathrm{do}(\{i\},\xi_i)}}\big(X_j \in \mathcal{B}_j\big) ,
\end{aligned}$$

where we used $X_i \perp\!\!\!\perp \boldsymbol{E}_{\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i})}$ in the fourth equality.

If we assume $i \notin \widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \setminus \mathcal{A}_j^{\setminus i}$ instead of $i \in \mathrm{pa}(\mathcal{A}_j^{\setminus i}) \setminus \mathcal{A}_j^{\setminus i}$, then we similarly arrive at the same conclusion. □

Proof of Corollary 7.3.2. This follows directly from Proposition 7.1.1 and 7.2.2. □

APPENDIX D: MEASURABLE SELECTION THEOREMS

In this section we derive some lemmas and state two measurable selection theorems that are used in several proofs in Appendix C. First we introduce the measure theoretic notation and terminology needed to understand the results (see (Kechris, 1995) for more details).

Definition D.1 (Standard measurable space). *A measurable space* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ *is a* standard measurable space *if it is isomorphic to* $(\boldsymbol{\mathcal{Y}}, \mathcal{B}(\boldsymbol{\mathcal{Y}}))$*, where* $\boldsymbol{\mathcal{Y}}$ *is a Polish space, i.e., a separable completely* *metrizable space,*[21] *and* $\mathcal{B}(\boldsymbol{\mathcal{Y}})$ *are the Borel subsets of* $\boldsymbol{\mathcal{Y}}$*, i.e., the* σ*-algebra generated by the open sets in* $\boldsymbol{\mathcal{Y}}$*. A measure space* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \boldsymbol{\mu})$ *is a* standard probability space *if* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ *is a standard measurable space and* $\boldsymbol{\mu}$ *is a probability measure.*

Examples of standard measurable spaces are the open and closed subsets of $\mathbb{R}^d$, and the finite sets with the usual complete metric. If we say that $\boldsymbol{\mathcal{X}}$ is a standard measurable space, then we implicitly assume that there exists a σ-algebra $\boldsymbol{\Sigma}$ such that $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ is a standard measurable space. Similarly, if we say that $\boldsymbol{\mathcal{X}}$ is a standard probability space with probability measure $\mathbb{P}_{\boldsymbol{\mathcal{X}}}$, then we implicitly assume that there exists a σ-algebra $\boldsymbol{\Sigma}$ such that $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \mathbb{P}_{\boldsymbol{\mathcal{X}}})$ is a standard probability space.

DEFINITION D.2 (Analytic set). *Let* $\boldsymbol{\mathcal{X}}$ *be a Polish space. A set* $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{X}}$ *is called* analytic *if there exist a Polish space* $\boldsymbol{\mathcal{Y}}$ *and a continuous mapping* $\boldsymbol{f} : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$ *with* $\boldsymbol{f}(\boldsymbol{\mathcal{Y}}) = \boldsymbol{\mathcal{A}}$*.*

LEMMA D.3. *Given a standard measurable space* $\boldsymbol{\mathcal{X}}$ *and* $\boldsymbol{\mathcal{Y}}$ *and a measurable mapping* $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$*, then*

1. *every measurable set* $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{X}}$ *is analytic;*
2. *if the subsets* $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{X}}$ *and* $\tilde{\boldsymbol{\mathcal{A}}} \subseteq \boldsymbol{\mathcal{Y}}$ *are analytic, then the sets* $\boldsymbol{f}(\boldsymbol{\mathcal{A}})$ *and* $\boldsymbol{f}^{-1}(\tilde{\boldsymbol{\mathcal{A}}})$ *are analytic.*

PROOF. From Proposition 13.7 in (Kechris, 1995) it follows that every measurable set $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{X}}$ is analytic. From Proposition 14.4.(ii) in (Kechris, 1995) it follows that the image and the preimage of an analytic set is an analytic set. □

DEFINITION D.4 ($\boldsymbol{\mu}$-measurability). *Consider a measure space* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma}, \boldsymbol{\mu})$*. A set* $\boldsymbol{\mathcal{E}} \subseteq \boldsymbol{\mathcal{X}}$ *is called a* $\boldsymbol{\mu}$-null set *if there exists a* $\boldsymbol{\mathcal{A}} \in \boldsymbol{\Sigma}$ *with* $\boldsymbol{\mathcal{E}} \subseteq \boldsymbol{\mathcal{A}}$ *and* $\boldsymbol{\mu}(\boldsymbol{\mathcal{A}}) = 0$*. We denote the class of* $\boldsymbol{\mu}$*-null sets by* $\boldsymbol{\mathcal{N}}$*, and we denote the* σ*-algebra generated by* $\boldsymbol{\Sigma} \cup \boldsymbol{\mathcal{N}}$ *by* $\bar{\boldsymbol{\Sigma}}$*, and its members are called the* $\boldsymbol{\mu}$-measurable *sets. Note that each member of* $\bar{\boldsymbol{\Sigma}}$ *is of the form* $\boldsymbol{\mathcal{A}} \cup \boldsymbol{\mathcal{E}}$ *with* $\boldsymbol{\mathcal{A}} \in \boldsymbol{\Sigma}$ *and* $\boldsymbol{\mathcal{E}} \in \boldsymbol{\mathcal{N}}$*. The measure* $\boldsymbol{\mu}$ *is extended to a measure* $\bar{\boldsymbol{\mu}}$ *on* $\bar{\boldsymbol{\Sigma}}$*, by* $\bar{\boldsymbol{\mu}}(\boldsymbol{\mathcal{A}} \cup \boldsymbol{\mathcal{E}}) = \boldsymbol{\mu}(\boldsymbol{\mathcal{A}})$ *for every* $\boldsymbol{\mathcal{A}} \in \boldsymbol{\Sigma}$ *and* $\boldsymbol{\mathcal{E}} \in \boldsymbol{\mathcal{N}}$*, and is called its* completion*. A mapping* $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ *between measurable spaces is called* $\boldsymbol{\mu}$-measurable *if the inverse image* $\boldsymbol{f}^{-1}(\boldsymbol{\mathcal{C}})$ *of every measurable set* $\boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{Y}}$ *is* $\boldsymbol{\mu}$*-measurable.*

DEFINITION D.5 (Universal measurability). *Consider a standard measurable space* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$*. A set* $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{X}}$ *is called* universally measurable *if it is* $\boldsymbol{\mu}$*-measurable for every* σ*-finite measure*[22] $\boldsymbol{\mu}$ *on* $\boldsymbol{\mathcal{X}}$ *(i.e., in particular every probability measure). A mapping* $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ *between standard measurable spaces is* universally measurable *if it is* $\boldsymbol{\mu}$*-measurable for every* σ*-finite measure* $\boldsymbol{\mu}$*.*

[21] A *metrizable space* is a topological space $\boldsymbol{\mathcal{X}}$ for which there exists a metric d such that $(\boldsymbol{\mathcal{X}}, d)$ is a metric space and induces the topology on $\boldsymbol{\mathcal{X}}$. For a metric space $(\boldsymbol{\mathcal{X}}, d)$, a *Cauchy sequence* is a sequence $(x_n)_{n\in\mathbb{N}}$ of elements of $\boldsymbol{\mathcal{X}}$ such that for every $\epsilon > 0$ there exists an $N \in \mathbb{N}$ such that for all natural numbers $p, q > N$ we have $d(x_n, x_m) < \epsilon$. We call $(\boldsymbol{\mathcal{X}}, d)$ *complete* if every Cauchy sequence has a limit in $\boldsymbol{\mathcal{X}}$. A *completely metrizable space* is a topological space $\boldsymbol{\mathcal{X}}$ for which there exists a metric d such that $(\boldsymbol{\mathcal{X}}, d)$ is a complete metric space that induces the topology on $\boldsymbol{\mathcal{X}}$. A topological space $\boldsymbol{\mathcal{X}}$ is called *separable* if it contains a countable dense subset, i.e., there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of elements in $\boldsymbol{\mathcal{X}}$ such that every non-empty open subset of $\boldsymbol{\mathcal{X}}$ contains at least one element of the sequence. A separable completely metrizable space is called a *Polish space* (see (Cohn, 2013) and (Kechris, 1995) for more details).

[22] A measure $\boldsymbol{\mu}$ on a measurable space $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ is called σ*-finite* if $\boldsymbol{\mathcal{X}} = \cup_{n\in\mathbb{N}}\boldsymbol{\mathcal{A}}_n$, with $\boldsymbol{\mathcal{A}}_n \in \boldsymbol{\Sigma}$, $\boldsymbol{\mu}(\boldsymbol{\mathcal{A}}_n) < \infty$.

LEMMA D.6. *Given a standard probability space $\boldsymbol{\mathcal{E}}$ with probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ and an analytic set $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{E}}$. Then $\boldsymbol{\mathcal{A}}$ is $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measurable and there exist measurable sets $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{T}} \subseteq \boldsymbol{\mathcal{E}}$ such that $\boldsymbol{\mathcal{S}} \subseteq \boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{T}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}}) = \bar{\mathbb{P}}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{A}}) = \mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{T}})$, where $\bar{\mathbb{P}}_{\boldsymbol{\mathcal{E}}}$ is the completion of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$.*

PROOF. Let $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{E}}$ be an analytic set. Since every analytic set in a standard measurable space is a universally measurable set (see Theorem 21.10 in (Kechris, 1995)), we know that $\boldsymbol{\mathcal{A}}$ is a universally measurable set, and hence it is in particular a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measurable set. Thus, there exist a measurable set $\boldsymbol{\mathcal{S}} \subseteq \boldsymbol{\mathcal{E}}$ and a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set $\boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{E}}$ such that $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{S}} \cup \boldsymbol{\mathcal{C}}$ and $\bar{\mathbb{P}}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{A}}) = \mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}})$, where $\bar{\mathbb{P}}_{\boldsymbol{\mathcal{E}}}$ is the completion of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$. Moreover, there exists a measurable set $\tilde{\boldsymbol{\mathcal{C}}} \subseteq \boldsymbol{\mathcal{E}}$ such that $\boldsymbol{\mathcal{C}} \subseteq \tilde{\boldsymbol{\mathcal{C}}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\tilde{\boldsymbol{\mathcal{C}}}) = 0$. Let $\boldsymbol{\mathcal{T}} := \boldsymbol{\mathcal{S}} \cup \tilde{\boldsymbol{\mathcal{C}}}$, then $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\mathcal{T}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{T}}) = \mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}})$. □

LEMMA D.7. *Consider a $\boldsymbol{\mu}$-measurable mapping $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$. If $\boldsymbol{\mathcal{Y}}$ is countably generated, then there exists a measurable mapping $\boldsymbol{g} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})$ holds $\boldsymbol{\mu}$-a.e..*

PROOF. Let the σ-algebra of $\boldsymbol{\mathcal{Y}}$ be generated by the countable generating set $\{\boldsymbol{\mathcal{C}}_n\}_{n\in\mathbb{N}}$. The $\boldsymbol{\mu}$-measurable set $\boldsymbol{f}^{-1}(\boldsymbol{\mathcal{C}}_n) = \boldsymbol{\mathcal{A}}_n \cup \boldsymbol{\mathcal{E}}_n$ for some $\boldsymbol{\mathcal{A}}_n \in \boldsymbol{\Sigma}$ and some $\boldsymbol{\mathcal{E}}_n \in \boldsymbol{\mathcal{N}}$ and hence there is some $\boldsymbol{\mathcal{E}}_n \subseteq \boldsymbol{\mathcal{B}}_n \in \boldsymbol{\Sigma}$ such that $\boldsymbol{\mu}(\boldsymbol{\mathcal{B}}_n) = 0$. Let $\hat{\boldsymbol{\mathcal{B}}} = \cup_{n\in\mathbb{N}}\boldsymbol{\mathcal{B}}_n$, $\hat{\boldsymbol{\mathcal{A}}}_n = \boldsymbol{\mathcal{A}}_n \setminus \hat{\boldsymbol{\mathcal{B}}}$ and $\hat{\boldsymbol{\mathcal{A}}} = \cup_{n\in\mathbb{N}}\hat{\boldsymbol{\mathcal{A}}}_n$, then $\boldsymbol{\mu}(\hat{\boldsymbol{\mathcal{B}}}) = 0$, $\hat{\boldsymbol{\mathcal{A}}}$ and $\hat{\boldsymbol{\mathcal{B}}}$ are disjoint and $\boldsymbol{\mathcal{X}} = \hat{\boldsymbol{\mathcal{A}}} \cup \hat{\boldsymbol{\mathcal{B}}}$. Now define the mapping $\boldsymbol{g} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ by

$$\boldsymbol{g}(\boldsymbol{x}) := \begin{cases} \boldsymbol{f}(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in \hat{\boldsymbol{\mathcal{A}}}, \\ \boldsymbol{y}_0 & \text{otherwise,} \end{cases}$$

where for $\boldsymbol{y}_0$ we can take an arbitrary point in $\boldsymbol{\mathcal{Y}}$. This mapping $\boldsymbol{g}$ is measurable since for each generator $\boldsymbol{\mathcal{C}}_n$ we have

$$\boldsymbol{g}^{-1}(\boldsymbol{\mathcal{C}}_n) = \begin{cases} \hat{\boldsymbol{\mathcal{A}}}_n & \text{if } \boldsymbol{y}_0 \notin \boldsymbol{\mathcal{C}}_n, \\ \hat{\boldsymbol{\mathcal{A}}}_n \cup \hat{\boldsymbol{\mathcal{B}}} & \text{otherwise.} \end{cases}$$

is in $\boldsymbol{\Sigma}$. Moreover, $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x})$ $\boldsymbol{\mu}$-almost everywhere. □

With this result at hand we can now prove the first measurable selection theorem.

THEOREM D.8 (Measurable selection theorem). *Given a standard probability space $\boldsymbol{\mathcal{E}}$ with probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$, a standard measurable space $\boldsymbol{\mathcal{X}}$ and a measurable set $\boldsymbol{\mathcal{S}} \subseteq \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{\mathcal{E}} \setminus \boldsymbol{pr}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}})$ is a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set, where $\boldsymbol{pr}_{\boldsymbol{\mathcal{E}}} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{E}}$ is the projection mapping on $\boldsymbol{\mathcal{E}}$. Then there exists a measurable mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ such that $(\boldsymbol{e}, \boldsymbol{g}(\boldsymbol{e})) \in \boldsymbol{\mathcal{S}}$ for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$.*

PROOF. Take the subset $\hat{\boldsymbol{\mathcal{E}}} := \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{B}}$, for some measurable set $\boldsymbol{\mathcal{B}} \supseteq \boldsymbol{\mathcal{E}} \setminus \boldsymbol{pr}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}})$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{B}}) = 0$, and note that $\hat{\boldsymbol{\mathcal{E}}}$ is a standard measurable space (see Corollary 13.4 in (Kechris, 1995)) and $\hat{\boldsymbol{\mathcal{E}}} \subseteq \boldsymbol{pr}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{S}})$. Let $\hat{\boldsymbol{\mathcal{S}}} = \boldsymbol{\mathcal{S}} \cap (\hat{\boldsymbol{\mathcal{E}}} \times \boldsymbol{\mathcal{X}})$. Because the set $\hat{\boldsymbol{\mathcal{S}}}$ is measurable, it is in particular analytic (see Lemma D.3). It follows by the Jankov-von Neumann Theorem (see Theorem 18.8 or 29.9 in (Kechris, 1995)) that $\hat{\boldsymbol{\mathcal{S}}}$ has a universally measurable uniformizing function, that is, there exists a universally measurable mapping $\hat{\boldsymbol{g}} : \hat{\boldsymbol{\mathcal{E}}} \to \boldsymbol{\mathcal{X}}$ such that for all $\boldsymbol{e} \in \hat{\boldsymbol{\mathcal{E}}}$, $(\boldsymbol{e}, \hat{\boldsymbol{g}}(\boldsymbol{e})) \in \hat{\boldsymbol{\mathcal{S}}}$. Hence, in particular, it is $\mathbb{P}_{\boldsymbol{\mathcal{E}}}\big|_{\hat{\boldsymbol{\mathcal{E}}}}$-measurable, where $\mathbb{P}_{\boldsymbol{\mathcal{E}}}\big|_{\hat{\boldsymbol{\mathcal{E}}}}$ is the restriction of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ to $\hat{\boldsymbol{\mathcal{E}}}$.

Now define the mapping $\boldsymbol{g}^* : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ by

$$\boldsymbol{g}^*(\boldsymbol{e}) := \begin{cases} \hat{\boldsymbol{g}}(\boldsymbol{e}) & \text{if } \boldsymbol{e} \in \hat{\boldsymbol{\mathcal{E}}} \\ \boldsymbol{x}_0 & \text{otherwise,} \end{cases}$$

where for $\boldsymbol{x}_0$ we can take an arbitrary point in $\boldsymbol{\mathcal{X}}$. Then this mapping $\boldsymbol{g}^*$ is $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measurable. To see this, take any measurable set $\boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{X}}$, then

$$\boldsymbol{g}^{*-1}(\boldsymbol{\mathcal{C}}) = \begin{cases} \hat{\boldsymbol{g}}^{-1}(\boldsymbol{\mathcal{C}}) & \text{if } \boldsymbol{x_0} \notin \boldsymbol{\mathcal{C}} \\ \hat{\boldsymbol{g}}^{-1}(\boldsymbol{\mathcal{C}}) \cup \boldsymbol{\mathcal{B}} & \text{otherwise.} \end{cases}$$

Because $\hat{\boldsymbol{g}}^{-1}(\boldsymbol{\mathcal{C}})$ is $\mathbb{P}_{\boldsymbol{\mathcal{E}}}\big|_{\hat{\boldsymbol{\mathcal{E}}}}$-measurable it is also $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measurable and thus $\boldsymbol{g}^{*-1}(\boldsymbol{\mathcal{C}})$ is $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measurable.

By Lemma D.7 and the fact that standard measurable spaces are countably generated (see Proposition 12.1 in (Kechris, 1995)), we prove the existence of a measurable mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{g}^* = \boldsymbol{g}$ $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.e. and thus it satisfies $(\boldsymbol{e}, \boldsymbol{g}(\boldsymbol{e})) \in \boldsymbol{\mathcal{S}}$ for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$. □

This theorem rests on the assumption that the standard measurable space $\boldsymbol{\mathcal{E}}$ has a probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$. If this space becomes the product space $\boldsymbol{\mathcal{Y}} \times \boldsymbol{\mathcal{E}}$, for some standard measurable space $\boldsymbol{\mathcal{Y}}$ where only the space $\boldsymbol{\mathcal{E}}$ has a probability measure, then in general this theorem does not hold anymore. However, if we assume in addition that the fibers of $\boldsymbol{\mathcal{S}}$ in $\boldsymbol{\mathcal{Y}}$ are σ-compact for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, then we can prove a second measurable selection theorem. A topological space is σ-compact if it is the union of countably many compact subspaces. For example, all countable discrete spaces, every interval of the real line, and moreover all the Euclidean spaces are σ-compact spaces.

THEOREM D.9 (Second measurable selection theorem). *Given a standard probability space $\boldsymbol{\mathcal{E}}$ with probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$, standard measurable spaces $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ and a measurable set $\boldsymbol{\mathcal{S}} \subseteq \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{Y}}$ such that $\boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{K}}_\sigma$ is a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set, where*

$$\boldsymbol{\mathcal{K}}_\sigma := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} \,:\, \forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{S}}_{(\boldsymbol{x},\boldsymbol{e})} \textit{ is non-empty and } \sigma\textit{-compact})\}\,,$$

with $\boldsymbol{\mathcal{S}}_{(\boldsymbol{x},\boldsymbol{e})}$ denoting the fiber over $(\boldsymbol{x}, \boldsymbol{e})$, that is

$$\boldsymbol{\mathcal{S}}_{(\boldsymbol{x},\boldsymbol{e})} := \{\boldsymbol{y} \in \boldsymbol{\mathcal{Y}} \,:\, (\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{y}) \in \boldsymbol{\mathcal{S}}\}\,.$$

Then there exists a measurable mapping $\boldsymbol{g} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Y}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ we have $(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{e})) \in \boldsymbol{\mathcal{S}}$.

PROOF. Take the subset $\hat{\boldsymbol{\mathcal{E}}} := \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{B}}$, for some measurable set $\boldsymbol{\mathcal{B}} \supseteq \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{K}}_\sigma$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{B}}) = 0$. Note that $\hat{\boldsymbol{\mathcal{E}}}$ is a standard measurable space, $\hat{\boldsymbol{\mathcal{E}}} \subseteq \boldsymbol{\mathcal{K}}_\sigma$ and $\hat{\boldsymbol{\mathcal{S}}} = \boldsymbol{\mathcal{S}} \cap (\boldsymbol{\mathcal{X}} \times \hat{\boldsymbol{\mathcal{E}}} \times \boldsymbol{\mathcal{Y}})$ is measurable. By assumption, for each $(\boldsymbol{x}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}} \times \hat{\boldsymbol{\mathcal{E}}}$ the fiber $\hat{\boldsymbol{\mathcal{S}}}_{(\boldsymbol{x},\boldsymbol{e})}$ is non-empty and σ-compact and hence by applying the Theorem of Arsenin-Kunugui (see Theorem 35.46 in (Kechris, 1995)) it follows that the set $\hat{\boldsymbol{\mathcal{S}}}$ has a measurable uniformizing function, that is, there exists a measurable mapping $\hat{\boldsymbol{g}} : \boldsymbol{\mathcal{X}} \times \hat{\boldsymbol{\mathcal{E}}} \to \boldsymbol{\mathcal{Y}}$ such that for all $(\boldsymbol{x}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}} \times \hat{\boldsymbol{\mathcal{E}}}$, $(\boldsymbol{x}, \boldsymbol{e}, \hat{\boldsymbol{g}}(\boldsymbol{x}, \boldsymbol{e})) \in \hat{\boldsymbol{\mathcal{S}}}$. Now define the mapping $\boldsymbol{g} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Y}}$ by

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{e}) := \begin{cases} \hat{\boldsymbol{g}}(\boldsymbol{x}, \boldsymbol{e}) & \text{if } \boldsymbol{e} \in \hat{\boldsymbol{\mathcal{E}}} \\ \boldsymbol{y}_0 & \text{otherwise,} \end{cases}$$

where for $\boldsymbol{y}_0$ we can take an arbitrary point in $\boldsymbol{\mathcal{Y}}$. This mapping $\boldsymbol{g}$ inherits the measurability from $\hat{\boldsymbol{g}}$ and it satisfies for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ that $(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{e})) \in \boldsymbol{\mathcal{S}}$. □

The next two lemmas provide some useful properties for the "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" quantifier.

LEMMA D.10. *Let $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$ be a measurable map between two standard measurable spaces. Let $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ be a probability measure on $\boldsymbol{\mathcal{E}}$ and let $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}} \circ \boldsymbol{\phi}^{-1}$ be its push-forward under $\boldsymbol{\phi}$. Let $\tilde{P} : \tilde{\boldsymbol{\mathcal{E}}} \to \{0,1\}$ be a property, i.e., a (measurable) boolean-valued function on $\tilde{\boldsymbol{\mathcal{E}}}$. Then the property $P = \tilde{P} \circ \boldsymbol{\phi}$ on $\boldsymbol{\mathcal{E}}$ holds $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.e. if and only if the property $\tilde{P}$ holds $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-a.e..*

PROOF. Assume the property $P = \tilde{P} \circ \boldsymbol{\phi}$ holds $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.e., then $\boldsymbol{\mathcal{C}} = \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : P(\boldsymbol{e}) = 1\}$ contains a measurable set $\boldsymbol{\mathcal{C}}^*$ with $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measure 1, i.e., $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\mathcal{C}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{C}}^*) = 1$. By Lemma D.3, $\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*)$ is analytic. By Lemma D.6, there exist measurable sets $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}$ such that $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*) \subseteq \boldsymbol{\mathcal{B}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{A}}) = \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{B}})$. Because $\boldsymbol{\phi}$ is measurable, $\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{A}})$ and $\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{B}})$ are both measurable. Also, $\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{A}}) \subseteq \boldsymbol{\phi}^{-1}(\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*)) \subseteq \boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{B}})$. As $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\phi}^{-1}(\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*))$, we must have that $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{B}})) \geq \mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{C}}^*) = 1$. Hence $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{A}}) = \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{B}}) = 1$. Note that as $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\mathcal{C}}$, $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*) \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}) \subseteq \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}}) = 1\}$. Hence the set $\tilde{\boldsymbol{\mathcal{C}}} := \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}}) = 1\}$ contains a measurable set of $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-measure 1, in other words, $\tilde{P}$ holds $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-a.s..

The converse is easier to prove. Suppose $\tilde{\boldsymbol{\mathcal{C}}} = \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}}) = 1\}$ contains a measurable set $\tilde{\boldsymbol{\mathcal{C}}}^*$ with $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-measure 1, i.e., $\tilde{\boldsymbol{\mathcal{C}}}^* \subseteq \tilde{\boldsymbol{\mathcal{C}}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\tilde{\boldsymbol{\mathcal{C}}}^*) = 1$. Because $\boldsymbol{\phi}$ is measurable, the set $\boldsymbol{\phi}^{-1}(\tilde{\boldsymbol{\mathcal{C}}}^*)$ is measurable and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\phi}^{-1}(\tilde{\boldsymbol{\mathcal{C}}}^*)) = 1$, and furthermore, $\boldsymbol{\phi}^{-1}(\tilde{\boldsymbol{\mathcal{C}}}^*) \subseteq \boldsymbol{\phi}^{-1}(\tilde{\boldsymbol{\mathcal{C}}}) = \boldsymbol{\mathcal{C}}$. □

LEMMA D.11 (Some properties for the for-almost-every quantifier). *Let $\boldsymbol{\mathcal{X}} = \mathcal{X} \times \tilde{\mathcal{X}}$ and $\boldsymbol{\mathcal{E}} = \mathcal{E} \times \tilde{\mathcal{E}}$ be products of non-empty standard measurable spaces and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}_{\mathcal{E}} \times \mathbb{P}_{\tilde{\mathcal{E}}}$ be the product measure of probability measures $\mathbb{P}_{\mathcal{E}}$ and $\mathbb{P}_{\tilde{\mathcal{E}}}$ on $\mathcal{E}$ and $\tilde{\mathcal{E}}$ respectively. Denote by "$\forall\!\!\!\!\forall \boldsymbol{e}$" the quantifier "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" and by "$\forall \boldsymbol{x}$" the quantifier "for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$", and similarly for their components, e.g. "$\forall\!\!\!\!\forall e$" for "for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$" and "$\forall x$" for "for all $x \in \mathcal{X}$". Then we have the following properties:*

1. $\forall\!\!\!\!\forall e : P(e) \implies \exists e : P(e)$ *(similarly to* $\forall x : P(x) \implies \exists x : P(x)$*);*
2. $\forall\!\!\!\!\forall e : P(e) \iff \forall\!\!\!\!\forall \boldsymbol{e} : P(e)$ *(similarly to* $\forall x : P(x) \iff \forall \boldsymbol{x} : P(x)$*);*
3. $\exists x \forall\!\!\!\!\forall e : P(x,e) \implies \forall\!\!\!\!\forall e \exists x : P(x,e)$ *(similarly to* $\exists x \forall e : P(x,e) \implies \forall e \exists x : P(x,e)$*);*
4. $\forall\!\!\!\!\forall e \forall x : P(x,e) \implies \forall x \forall\!\!\!\!\forall e : P(x,e)$ *(similarly to* $\forall e \forall x : P(x,e) \implies \forall x \forall e : P(x,e)$*);*
5. $\forall\!\!\!\!\forall \boldsymbol{e} : P(\boldsymbol{e}) \implies \exists \tilde{e} \forall\!\!\!\!\forall e : P(\boldsymbol{e})$ *(similarly to* $\forall \boldsymbol{x} : P(\boldsymbol{x}) \implies \exists \tilde{x} \forall x : P(\boldsymbol{x})$*);*
6. $\forall\!\!\!\!\forall e \forall x : P(x,e) \iff \forall\!\!\!\!\forall \boldsymbol{e} \forall \boldsymbol{x} : P(x,e)$*;*
7. $\forall\!\!\!\!\forall \boldsymbol{e} \forall \boldsymbol{x} : P(\boldsymbol{x}, \boldsymbol{e}) \implies \exists \tilde{e} \exists \tilde{x} \forall\!\!\!\!\forall e \forall x : P(\boldsymbol{x}, \boldsymbol{e})$*,*

where P denotes a property, i.e., a measurable boolean-valued function, on the corresponding measurable spaces and we write $\boldsymbol{e}$ and $\boldsymbol{x}$ for $(e, \tilde{e})$ and $(x, \tilde{x})$ respectively.

PROOF. We only prove the statements that may not be immediately obvious.

Property 2. Let $pr_{\mathcal{E}} : \boldsymbol{\mathcal{E}} \to \mathcal{E}$ be the projection mapping on $\mathcal{E}$. Then by Lemma D.10 we have

$$\forall\!\!\!\!\forall e : P(e) \iff \forall\!\!\!\!\forall \boldsymbol{e} : P \circ \mathrm{pr}_{\mathcal{E}}(\boldsymbol{e}) \iff \forall\!\!\!\!\forall \boldsymbol{e} : P(e)\,.$$

Property 4: We have

$$\begin{aligned}
&\forall\!\!\!\!\forall e \forall x : P(x,e) \\
&\implies \exists\, \mathbb{P}_{\mathcal{E}}\text{-null set } N \,\forall e \in \mathcal{E} \setminus N \,\forall x : P(x,e) \\
&\implies \exists\, \mathbb{P}_{\mathcal{E}}\text{-null set } N \,\forall x \,\forall e \in \mathcal{E} \setminus N : P(x,e) \\
&\implies \forall x \,\exists\, \mathbb{P}_{\mathcal{E}}\text{-null set } N \,\forall e \in \mathcal{E} \setminus N : P(x,e) \\
&\implies \forall x \forall\!\!\!\!\forall e : P(x,e)\,.
\end{aligned}$$

Property 5: Let $\boldsymbol{N}$ be a measurable $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set such that $P(\boldsymbol{e})$ holds for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}} \setminus \boldsymbol{N}$. Define for $\tilde{e} \in \tilde{\mathcal{E}}$ the set $N_{\tilde{e}} := \{e \in \mathcal{E} : (e, \tilde{e}) \in \boldsymbol{N}\}$. Note that the sets $N_{\tilde{e}}$ are measurable. From Fubini's theorem it follows that for $\mathbb{P}_{\tilde{\mathcal{E}}}$-almost every $\tilde{e} \in \tilde{\mathcal{E}}$ we have $\mathbb{P}_{\mathcal{E}}(N_{\tilde{e}}) = 0$. That is, there exists a measurable $\mathbb{P}_{\tilde{\mathcal{E}}}$-null set $\tilde{N}$ such that $\mathbb{P}_{\mathcal{E}}(N_{\tilde{e}}) = 0$ for all $\tilde{e} \in \tilde{\mathcal{E}} \setminus \tilde{N}$. Hence, there exists $\tilde{e} \in \tilde{\mathcal{E}} \setminus \tilde{N}$ such that $\mathbb{P}_{\mathcal{E}}(N_{\tilde{e}}) = 0$; for all $e \in \mathcal{E} \setminus N_{\tilde{e}}$, $P(\boldsymbol{e})$ then holds. This means $\exists \tilde{e} \forall e : P(\boldsymbol{e})$.

Property 7: We have

$$\begin{aligned} \forall \boldsymbol{e} \forall \boldsymbol{x} : P(\boldsymbol{x}, \boldsymbol{e}) &\implies \exists \tilde{e} \forall e \forall \boldsymbol{x} : P(\boldsymbol{x}, \boldsymbol{e}) \implies \exists \tilde{e} \forall e \forall \tilde{x} \forall x : P(\boldsymbol{x}, \boldsymbol{e}) \\ &\implies \exists \tilde{e} \forall \tilde{x} \forall e \forall x : P(\boldsymbol{x}, \boldsymbol{e}) \implies \exists \tilde{e} \exists \tilde{x} \forall e \forall x : P(\boldsymbol{x}, \boldsymbol{e}) , \end{aligned}$$

where in the first equivalence we used Property 5, in the third equivalence we used Property 4 and in the last equivalence we used Property 1. □

ACKNOWLEDGMENTS

We thank Robin Evans for some helpful discussion and Noud de Kroon, Tineke Blom and Alexander Ly for providing helpful comments on earlier drafts.

REFERENCES

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.

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.

Blom, T., Bongers, S. and Mooij, J. M. (2019). Beyond Structural Causal Models: Causal Constraints Models. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI-19)* (R. P. Adams and V. Gogate, eds.). AUAI Press.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, New York, USA.

Bongers, S. and Mooij, J. M. (2018). From Random Differential Equations to Structural Causal Models: the stochastic case. *arXiv.org preprint* arXiv:1803.08784 [cs.AI].

Bühlmann, P., Peters, J. and Ernest, J. (2014). CAM: Causal Additive Models, high-dimensional order search and penalized regression. *The Annals of Statistics* **42** 2526–2556.

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

Cohn, D. L. (2013). *Measure Theory*, 2nd ed. Birkhäuser, Boston, USA.

Cooper, G. F. (1997). A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships. *Data Mining and Knowledge Discovery* **1** 203–224.

Dawid, A. P. (2002). Influence Diagrams for Causal Modelling and Inference. *International Statistical Review* **70** 161–189.

Duncan, O. D. (1975). *Introduction to Structural Equation Models.* Academic Press, New York.

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.

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.

Evans, R. J. (2016). Graphs for Margins of Bayesian Networks. *Scandinavian Journal of Statistics* **43** 625–648.

Evans, R. J. (2018). Margins of discrete Bayesian networks. *The Annals of Statistics* **46** 2623–2656.

Forré, P. and Mooij, J. M. (2017). Markov Properties for Graphical Models with Cycles and Latent Variables. *arXiv.org preprint* arXiv:1710.08775 [math.ST].

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 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)* (A. Globerson and R. Silva, eds.). AUAI Press.

Forré, P. and Mooij, J. M. (2019). Causal Calculus in the Presence of Cycles, Latent Confounders and Selection Bias. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI-19)* (R. P. Adams and V. Gogate, eds.). AUAI Press.

Foygel, R., Draisma, J. and Drton, M. (2012). Half-trek Criterion for Generic Identifiability of Linear Structural Equation Models. *The Annals of Statistics* **40** 1682–1713.

Geiger, D. (1990). Graphoids: A Qualitative Framework for Probabilistic Inference Technical Report No. R-142, Computer Science Department, University of California, Los Angeles, USA.

Goldberger, A. S. and Duncan, O. D. (1973). *Structural Equation Models in the Social Sciences.* Seminar Press, New York.

Golub, G. and Kahan, W. (1965). Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* **2** 205–224.

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica* **11** 1–12.

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.

Hyttinen, A., Eberhardt, F. and Hoyer, P. O. (2012). Learning Linear Cyclic Causal Models with Latent Variables. *Journal of Machine Learning Research* **13** 3387–3439.

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, Oregon, USA.

Kechris, A. S. (1995). *Classical Descriptive Set Theory. Graduate Texts in Mathematics* **156**. Springer-Verlag, New York, USA.

Koster, J. T. A. (1996). Markov Properties of Nonrecursive Causal Models. *The Annals of Statistics* **24** 2148–2177.

Koster, J. T. A. (1999). On the Validity of the Markov Interpretation of Path Diagrams of Gaussian Structural Equations Systems with Correlated Errors. *Scandinavian Journal of Statistics* **26** 413–431.

Lauritzen, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Clarendon Press, Oxford.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H. G. (1990). Independence Properties of Directed Markov Fields. *Networks* **20** 491–505.

Lewis, D. K. (1979). Counterfactual Dependence and Time's Arrow. *Noûs* **13** 455–476.

Maathuis, M. H., Colombo, D., Kalisch, M. and Bühlmann, P. (2009). Estimating High-Dimensional Intervention Effects from Observational Data. *The Annals of Statistics* **37** 3133–3164.

Mani, S. (2006). A Bayesian Local Causal Discovery Framework, PhD thesis, University of Pittsburg.

Mason, S. J. (1953). Feedback Theory - Some Properties of Signal Flow Graphs. In *Proceedings of the IRE* **41** 1144-1156. IEEE.

Mason, S. J. (1956). Feedback Theory - Further Properties of Signal Flow Graphs. In *Proceedings of the IRE* **44** 920–926. IEEE.

Mooij, J. M. and Claassen, T. (2020). Constraint-Based Causal Discovery In The Presence Of Cycles. *arXiv.org preprint* arXiv:2005.00610 [math.ST].

Mooij, J. M. and Heskes, T. (2013). Cyclic Causal Discovery from Continuous Equilibrium Data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-13)* (A. Nicholson and P. Smyth, eds.) 431–439. AUAI Press, Corvallis, Oregon, USA.

Mooij, J. M., Janzing, D. and Schölkopf, B. (2013). From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-13)* (A. Nicholson and P. Smyth, eds.) 440–448. AUAI Press.

Mooij, J. M., Magliacane, S. and Claassen, T. (2019). Joint Causal Inference from Multiple Contexts. *arXiv.org preprint* arXiv:1611.10351 [cs.LG].

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. and Schölkopf, B. (2016). Distinguishing Cause from Effect using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research* **17** 1–102.

Neal, R. M. (2000). On Deducing Conditional Independence from *d*-Separation in Causal Graphs with Feedback. *Journal of Artificial Intelligence Research* **12** 87–91.

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, Oregon, USA.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, New York, USA.

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.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*, 1st ed. Basic Books, New York, USA.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society* **51** 406–413.

Peters, J., Janzing, D. and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* MIT Press, Cambridge, MA, USA.

Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2014). Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* **15** 2009–2053.

Pfister, N., Bauer, S. and Peters, J. (2019). Learning Stable and Predictive Structures in Kinetic Systems. *Proceedings of the National Academy of Sciences* **116** 25405–25411.

Richardson, T. S. (1996a). 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.

Richardson, T. S. (1996b). Discovering Cyclic Causal Structure Technical Report No. CMU-PHIL-68, Carnegie Mellon University.

Richardson, T. (2003). Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics* **30** 145–157.

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.

Richardson, T. S. and Robins, J. M. (2014). ACE Bounds; SEMs with Equilibrium Conditions. *Statistical Science* **29** 363-366.

Richardson, T. S. and Spirtes, P. (1999). Automated Discovery of Linear Feedback Models. In *Computation, Causation, and Discovery* (C. Glymour and G. F. Cooper, eds.) 253-304. MIT Press.

Richardson, T. S. and Spirtes, P. (2002). Ancestral Graph Markov Models. *The Annals of Statistics* **30** 962–1030.

Roese, N. J. (1997). Counterfactual Thinking. *Psychological Bulletin* **121** 133–148.

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 33rd Conference on Uncertainty in Artificial Intelligence (UAI-17)* (G. Elidan and K. Kersting, eds.). AUAI Press.

Rubenstein, P. K., Bongers, S., Schölkopf, B. and Mooij, J. M. (2018). From Deterministic ODEs to Dynamic Structural Causal Models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)* (A. Globerson and R. Silva, eds.). AUAI Press.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66** 688–701.

Shpitser, I. and Pearl, J. (2008). Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research* **9** 1941–1979.

Spirtes, P. (1993). Directed Cyclic Graphs, Conditional Independence, and Non-recursive Linear Structural Equation Models Technical Report No. CMU-PHIL-35, Carnegie Mellon University.

Spirtes, P. (1994). Conditional Independence in Directed Cyclic Graphical Models for Feedback Technical Report No. CMU-PHIL-54, Carnegie Mellon University.

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.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning.* MIT Press, Cambridge, Massachusetts.

Spirtes, P., Meek, C. and Richardson, T. S. (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.) 6, 211-252. The MIT Press.

Spirtes, P., Richardson, T., Meek, C., Scheines, R. and Glymour, C. (1998). Using Path Diagrams as a Structural Equation Modelling Tool. *Sociological Methods & Research* **27** 182–225.

Tian, J. (2002). Studies in Causal Reasoning and Learning Technical Report No. R-309, Cognitive Systems Laboratory, University of California, Los Angeles, USA.

Tian, J. and Pearl, J. (2001). Causal Discovery from Changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI-01)* (J. Breese and D. Koller, eds.) 512–521. Morgan Kaufmann, San Francisco, CA, USA.

Verma, T. S. (1993). Graphical Aspects of Causal Models Technical Report No. R-191, Computer Science Department, University of California, Los Angeles, USA.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research* **20** 557–585.

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.

Informatics Institute
University of Amsterdam
The Netherlands
E-mail: s.r.bongers@uva.nl
p.d.forre@uva.nl

Department of Mathematical Sciences
University of Copenhagen
Denmark
E-mail: jonas.peters@math.ku.dk

Empirical Inference
Max Planck Institute for Intelligent Systems
Germany
E-mail: bernhard.schoelkopf@tuebingen.mpg.de

Korteweg-De Vries Institute
University of Amsterdam
The Netherlands
E-mail: j.m.mooij@uva.nl