Structural Causal Models: Cycles, Marginalizations, Exogenous Reparametrizations and Reductions

Stephan Bongers, Jonas Peters, Bernhard Schölkopf, Joris M. Mooij

November 21, 2016

Abstract

Structural causal models (SCMs), also known as non-parametric structural equation models (NP-SEMs), are widely used for causal modeling purposes. In this paper, we give a rigorous treatment of structural causal models, dealing with measure-theoretic complications that arise in the presence of cyclic relations. The central question studied in this paper is: given a (possibly cyclic) SCM defined on a large system (consisting of observable endogenous and latent exogenous variables), can we "project it down" to an SCM that describes a subsystem (consisting of a subset of the observed endogenous variables and possibly different latent exogenous variables) in order to obtain a more parsimonious but equivalent representation of the subsystem? We define a *marginalization* operation that effectively removes a subset of the endogenous variables from the model, and a class of mappings, *exogenous reparameterizations*, that can be used to reduce the space of exogenous variables. We show that both operations preserve the causal semantics of the model and that under mild conditions they can lead to a significant reduction of the model complexity, at least in terms of the number of variables in the model. We argue that for the task of estimating an SCM from data, the existence of "smooth" reductions would be desirable. We provide several conditions under which the existence of such reductions can be shown, but also provide a counterexample that shows that such reductions do not exist in general. The latter result implies that existing approaches to estimate linear or Markovian SCMs from data cannot be extended to general SCMs.

1 Introduction

Structural causal models (SCMs), also known as (non-parametric) structural equation models (NP-SEMs), are widely used for causal modeling purposes (Pearl, 2009; Spirtes et al., 2000). In these models, the causal relationships are expressed in the form of deterministic, functional relationships, and probabilities are introduced through the assumption that certain variables are exogenous, 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 are closely related to causal Bayesian networks. These directed graphical models have relatively simple definitions in terms of various equivalent Markov properties, are computationally scalable and have several nice statistical properties (Pearl, 1988; Lauritzen et al., 1990; Lauritzen, 1996). However, an important limitation of acyclic SCMs is that they cannot represent systems that involve feedback. For that purpose, cyclic SCMs (or non-recursive SEMs) form an appropriate model class (see e.g., Mooij et al., 2013). In contrast to the acyclic case, however, cyclic SCMs have enjoyed less attention and are not as well understood as their acyclic counterparts, although some progress has been made in the case of discrete (Pearl and Dechter, 1996; Neal, 2000) and linear models (Spirtes, 1994, 1995; Hyttinen et al., 2012).

When allowing for cyclic relationships 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 cyclic case, however, this property may be violated, and consequently, the SCM is either ill-defined or allows for multiple probability distributions, which leads to ambiguity. Worse, 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. Furthermore, even if the functional relations of the SCM are measurable functions, and if the SCM is uniquely solvable, the solutions of the SCM may not be measurable. All these issues make SCMs in the cyclic setting a notoriously more complicated class of models to work with than the class of acyclic SCMs. In this paper we will give a general, rigorous treatment including both acyclic and cyclic SCMs, dealing with the technical complications that arise in the presence of cyclic relations.

Consider a (possibly cyclic) SCM that is defined on a large system, consisting of observable endogenous variables and latent exogenous variables. For example, such an SCM could be obtained by starting with an ordinary differential equation model and considering its equilibrium states (Mooij et al., 2013). Often, one is not interested in modeling the entire system, but would like to focus attention on a smaller subsystem instead, for example due to practical limitations on the measurability of some of the endogenous variables in the model. The central question of this paper is whether and how the SCM that represents the large system can be "projected down" to a more parsimonious SCM that represents the smaller subsystem.[1] The SCM representing the subsystem is defined on a subset of the original endogenous variables and may have different latent exogenous variables, but should preserve the causal semantics of the subsystem variables. We decompose this question into two parts, deal-

[1]This question relates to the concept of "latent projection" (Verma, 1993; Tian, 2002) that has been formulated for acyclic graphical models and to ancestral graph Markov models (Richardson and Spirtes, 2002).

ing with the endogenous and exogenous variables separately. We define a *marginalization* operation that only operates on the endogenous variables of an SCM, and *exogenous reparameterizations*, a class of mappings that operate only on the exogenous (latent) variables. By combining a marginalization with a suitable exogenous reparameterization, one may arrive at a more parsimonious causal model representing the subsystem. This is advantageous when estimating an SCM from data, as parsimonious models typically need less data for obtaining reliable estimates.

Marginalizing an SCM over a set of endogenous variables allows us to focus our causal modeling efforts on a part of a system and to ignore the rest of the system. On a high level, the concept is analogous to marginalizing a probability distribution over a set of variables: this operation also reduces the description of a set of variables to a description of a subset of those, but the difference is that there the description is purely probabilistic, whereas for SCMs it is causal and probabilistic. The idea that leads to the definition of marginalization for SCMs is to think of the subset of endogenous variables of interest as a subsystem that can interact with the rest of the system. Under some suitable conditions, 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. Marginalizing over this subset of variables effectively removes this subsystem from the model. We will show that this marginalization operation indeed preserves the causal semantics, i.e., all interventional distributions of the remaining endogenous variables induced by the original SCM are identical to those induced by its marginalization.

The marginalization operation maps an SCM to another SCM that has only a subset of the endogenous variables, while preserving the causal semantics. Similarly, one can map an SCM to another SCM with a possibly different set of exogenous variables and a different distribution on those variables, while preserving the causal semantics. This can be achieved via a class of mappings that we will call *exogenous reparametrizations*. They form an interesting class in their own right, including for example certain symmetries that act on the space of SCMs. Moreover, since marginalization acts only on the endogenous part of the system, this may still lead to very complex models, even for systems with only a few endogenous variables. In certain cases, more parsimonious representations can be obtained by using exogenous reparametrizations that reduce the (dimensionality of the) space of the exogenous variables.

The question of how to represent the influence of the surroundings of the system of interest in a more parsimonious way leads us to introduce the concept of *reduction*. A reduction of an SCM is an interventionally equivalent SCM (i.e., all interventional distributions induced by the original SCM are identical to those induced by its reduction) defined on the same endogenous variables but with a lower-dimensional space of exogenous variables. For certain classes of SCMs, such reductions can be obtained by performing a suitable exogenous reparametrization. This has presumably been known for a long time for linear SCMs,[2] but we are not aware

[2] The earliest work that makes use of such reductions, to the best of our knowledge, is Hyttinen et al. (2012).

of any general treatment for possibly nonlinear and cyclic SCMs. Interestingly, for SCMs with real-valued variables, it turns out that one can always find an exogenous reparametrization that reduces the exogenous variables down to a single one-dimensional real-valued exogenous variable. However, in general such heavily reduced SCMs should be impossible to estimate from data, as their causal mechanisms are typically very wild.

We observe that when starting with a linear SCM, a reduction exists that is also linear, and hence can be estimated relatively easily from data. Also, when starting with a Markovian SCM, reductions exist that will typically be at least as smooth as the original SCM. These observations suggest that "smooth" reductions, which are easier to estimate from data, might exist in general. However, we provide a counterexample that shows that this is actually not the case. This result may have implications for approaches to estimating SCMs from data.

In the next section, we will give a formal definition of SCMs, discuss their causal interpretation, and consider various equivalence relations between SCMs. In Section 3, we will introduce and study the marginalization operation. Section 4 introduces exogenous reparameterizations and studies their properties. This will be a useful tool for Section 5 that defines and studies the properties of reductions. We wrap up and discuss the implications of this work in the last section.

2 Structural causal models

In this section, we will start by formally defining structural causal models (SCMs) and their solutions. We discuss their causal structure and show that uniquely solvable SCMs give rise to a unique probability distribution on the variables. We will discuss interventions of SCMs, which give rise to their causal semantics. Finally, we will define various equivalence relations between SCMs.

2.1 Definition

One usually considers random variables to be part of an SCM. We will use a slightly different approach here that makes it easier to deal with cyclic SCMs. Our approach is to strip off the random variables from the definition of the SCM, which ensures that SCMs form a well-defined model class that is closed under any perfect intervention (see also Section 2.4). In Section 2.3, we will give conditions under which these SCMs induce a well-defined set of (measurable) random variables.

Definition 1 *A* structural causal model (SCM) *is a tuple*[3]

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

where

- $\mathcal{I}$ *is a finite index set of* endogenous variables*,*

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

- $\mathcal{J}$ *is a finite index set of* exogenous variables*,*
- $\boldsymbol{\mathcal{X}} = \prod_{i\in\mathcal{I}} \mathcal{X}_i$ *is the product of the domains of the endogenous variables where each domain $\mathcal{X}_i$ is a standard*[4] *measurable space,*
- $\boldsymbol{\mathcal{E}} = \prod_{j\in\mathcal{J}} \mathcal{E}_j$ *is the product of the domains of the exogenous variables where each domain $\mathcal{E}_j$ is a standard measurable space,*
- $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ *is a measurable function that specifies the* causal mechanisms*,*
- $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ *is a probability measure on $\boldsymbol{\mathcal{E}}$.*

We will denote the collection of all SCMs as **SCM**. Although it is common to assume the absence of causal feedback (i.e., cyclic causal relations, see Definition 8), we will make no such assumption here. When allowing for cycles, it turns out that the most natural setting is obtained when allowing for self-loops as well. Under certain conditions, these self-loops can be removed. We will discuss this later in more detail.

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

Definition 2 *For an SCM $\mathcal{M}$, we call the set of equations*

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}), \qquad \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}} \tag{1}$$

the structural equations *of the structural causal model $\mathcal{M}$.*

Somebody familiar with structural causal models will note that we are still missing an important ingredient: random variables. In our approach, they come in as follows:

Definition 3 *We call a pair of random variables $(\boldsymbol{E}, \boldsymbol{X})$ 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. *$\boldsymbol{E}$ takes values in $\boldsymbol{\mathcal{E}}$,*
2. *$\boldsymbol{X}$ takes values in $\boldsymbol{\mathcal{X}}$,*
3. *$\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$, i.e., the distribution of $\boldsymbol{E}$ is equal to $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$,*
4. *the* structural equations

$$\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$$

are satisfied almost surely.

Often, the endogenous random variables $\boldsymbol{X}$ can be observed, and the exogenous random variables $\boldsymbol{E}$ are latent. Latent exogenous variables are often referred to as "disturbance terms" or "noise variables". Note that for two exogenous variables $\boldsymbol{E}$ and $\tilde{\boldsymbol{E}}$ we have that if $\boldsymbol{E} = \tilde{\boldsymbol{E}}$ a.s., then $\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}^{\tilde{\boldsymbol{E}}}$.

[4] A standard measurable space is a measurable space (Ω, Σ) that is isomorphic to a measurable space $(\tilde{\Omega}, \tilde{\Sigma})$, where $\tilde{\Omega}$ is a Polish space (i.e., a complete separable metric space) and $\tilde{\Sigma} = \mathcal{B}(\tilde{\Omega})$ are the Borel subsets of $\tilde{\Omega}$ (i.e., the σ-algebra generated by the open sets in $\tilde{\Omega}$). In several proofs we will assume without loss of generality that the standard measurable space *is* a Polish space Ω with σ-algebra $\mathcal{B}(\Omega)$. Examples of standard measurable spaces are open and closed subsets of $\mathbb{R}^d$, and finite sets with the complete σ-algebra. See for example Cohn (2013) for more details.

The reason for stripping off the random variables from the definition of the structural causal model is that, as we will see in Section 2.4, it may happen that an SCM without any solution may have a solution after performing a perfect intervention on the SCM. Conversely, it may also happen that intervening on an SCM with a solution gives an SCM without solution.

In terms of the solutions of an SCM, this definition leads to an ambiguity, in the sense that we could have two different SCMs with exactly the same solutions. This ambiguity follows from the fact that the structural equations are only satisfied almost surely. To accommodate for this it seems natural to not differentiate between causal mechanisms that have different solutions on at most a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set of exogenous variables:

Definition 4 *Two causal mechanisms $f_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$ and $\tilde{f}_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$ are* equivalent*, denoted $f_i \equiv \tilde{f}_i$, if for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$:*[5]

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}} : x_i = f_i(\boldsymbol{x}, \boldsymbol{e}) \iff x_i = \tilde{f}_i(\boldsymbol{x}, \boldsymbol{e})$$

If $f_i \equiv \tilde{f}_i$ for all $i \in \mathcal{I}$, then we say that $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$ are equivalent, and write $\boldsymbol{f} \equiv \tilde{\boldsymbol{f}}$.

This gives rise to an equivalence relation for SCMs:

Definition 5 *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 $\mathcal{M} \equiv \tilde{\mathcal{M}}$) if all their causal mechanisms are equivalent, i.e., $f_i \equiv \tilde{f}_i$ for all $i \in \mathcal{I}$.*

This equivalence relation $\equiv$ on **SCM**, the set of all SCMs, gives rise to the quotient set of equivalence classes of SCMs, which will be denoted by $\mathbf{SCM}_{eq} := \mathbf{SCM}/\equiv$. Note that two equivalent SCMs only differ by their causal mechanism. Importantly, equivalent SCMs have the same solutions.

In this paper we will prove several properties and define several operations and relations on $\mathbf{SCM}_{eq}$. 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. In both cases we say that the property or operation *descends* to the set of equivalences classes.

2.2 Causal structure

The functional relations in an SCM, i.e., the causal mechanisms, encode a certain causal structure, that we can summarize by a directed graph. Before we define this graph, we first state some standard terminology on directed graphs.

A *directed graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where we have a set of nodes $\mathcal{V}$ and a set of directed edges $\mathcal{E}$, which is a subset $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ of ordered pairs

[5]Please note that in general the quantifier "for $\mathbb{P}$-almost every" does not commute with the quantifier "for all".

of nodes. Each element $(i,j) \in \mathcal{E}$ can be represented by the directed edge $i \to j$, in particular $(i,i) \in \mathcal{E}$ represents a self-loop $i \to i$. A *walk* between i and j in $\mathcal{G}$ is a sequence of edges $(\epsilon_1, \ldots, \epsilon_n)$ such that there exists a sequence of nodes $(i = i_0, i_1, \ldots, i_{n-1}, i_n = j)$ for some $n \geq 0$, such that either $\epsilon_k = (i_{k-1}, i_k)$ or $\epsilon_k = (i_k, i_{k-1})$ for $k = 1, 2, \ldots, n$ (note that $n = 0$ corresponds with a walk consisting of a single node); if all nodes $i_0, \ldots, i_n$ are distinct, it is called a *path*. A walk, or path, of the form $i \to \cdots \to j$, i.e., such that $\epsilon_k = (i_{k-1}, i_k)$ for all $k = 1, 2, \ldots, n$, is called a *directed walk, or path*, from i to j. A *cycle* is a sequence of edges $(\epsilon_1, \ldots, \epsilon_{n+1})$ such that $(\epsilon_1, \ldots, \epsilon_n)$ is a directed path from i to j and $\epsilon_{n+1} = (j, i)$. In particular, a self-loop is a cycle. Note a path cannot contain any cycles. A directed graph is *acyclic* if it contains no cycles, and is then referred to as a *directed acyclic graph (DAG)*.

For a node i we denote the set of *parents* of i by $\mathrm{pa}(i) = \{j \in \mathcal{V} : j \to i \text{ is a path in } \mathcal{G}\}$, the set of children of i by $\mathrm{ch}(i) = \{j \in \mathcal{V} : i \to j \text{ is a path in } \mathcal{G}\}$, the set of *ancestors* of i by $\mathrm{an}(i) = \{j \in \mathcal{V} : \text{there is a directed path from } j \text{ to } i \text{ in } \mathcal{G}\}$ and the set of *descendants* of i by $\mathrm{de}(i) = \{j \in \mathcal{V} : \text{there is a directed path from } i \text{ to } j \text{ in } \mathcal{G}\}$. We can apply these definitions to subsets $\mathcal{U} \subseteq \mathcal{V}$ by taking the union of these sets, for example $\mathrm{pa}(\mathcal{U}) = \cup_{i \in \mathcal{U}} \mathrm{pa}(i)$.

Definition 6 *Let $\mathcal{M}$ be an SCM. We say that*

- *(i)* *$i \in \mathcal{I}$ is a* direct cause *of $k \in \mathcal{I}$ if and only if there does not exist a measurable function*[6] *$\tilde{f}_k : \boldsymbol{\mathcal{X}}_{\setminus i} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_k$ such that $\tilde{f}_k \equiv f_k$ (See Definition 5).*
- *(ii)* *$j \in \mathcal{J}$ is a* direct cause *of $k \in \mathcal{I}$ if and only if there does not exist a measurable function $\tilde{f}_k : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}}_{\setminus j} \to \mathcal{X}_k$ such that $\tilde{f}_k \equiv f_k$.*

By definition no variable is a direct cause of an exogenous variable. Obviously, direct causal relations are preserved under the equivalence relation $\equiv$ on SCMs. Note that there always exists a representation $\tilde{\boldsymbol{f}}$ of the causal mechanism $\boldsymbol{f}$ such that each component depends only on its direct causes, i.e., each component $\tilde{f}_i$ can be considered without loss of generality to be a function $\tilde{f}_i : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(i) \cap \mathcal{I}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(i) \cap \mathcal{J}} \to \mathcal{X}_i$. Similarly, for example for $i \in \mathcal{I}$ the ancestral components of $\boldsymbol{f}_{\mathrm{an}(i)}$ of $\boldsymbol{f}$, one can always find a representation of $\boldsymbol{f}$ such that $\boldsymbol{f}_{\mathrm{an}(i)} : \boldsymbol{\mathcal{X}}_{\mathrm{an}(i) \cap \mathcal{I}} \times \boldsymbol{\mathcal{E}}_{\mathrm{an}(i) \cap \mathcal{J}} \to \boldsymbol{\mathcal{X}}_{\mathrm{an}(i) \cap \mathcal{I}}$ does not depend on non-ancestors of i.

The direct causal relations can be represented by a directed graph:

Definition 7 *Let $\mathcal{M}$ be an SCM. We define the* augmented graph $\mathcal{G}^a(\mathcal{M})$ *as the directed graph with nodes $\mathcal{I} \cup \mathcal{J}$ and directed edges $i \to j$ if and only if i is a direct cause of j.*

In particular, the augmented graph contains no directed edges between exogenous variables, because they are not functionally related by the causal mechanism, note this does not imply that the exogenous variables are jointly independent. This definition maps $\mathcal{M}$ to $\mathcal{G}^a(\mathcal{M})$, which we call the "augmented graph mapping" $\mathcal{G}^a$. By definition, $\mathcal{G}^a(\mathcal{M})$ is invariant under the equivalence of SCMs and hence the whole equivalence class of an SCM is mapped to a unique augmented graph.

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

Definition 8 *We call an SCM $\mathcal{M}$* acyclic *if $\mathcal{G}^a(\mathcal{M})$ is a DAG, otherwise we call $\mathcal{M}$* cyclic.

Most of the existing literature considers only acyclic SCMs.[7] A particular interesting class of acyclic SCMs are the, so called *Markovian* SCMs. A *Markovian* SCM is an SCM that is acyclic, that has jointly independent exogenous variables, i.e, exogenous probability distribution $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j\in\mathcal{J}} \mathbb{P}_{\boldsymbol{\mathcal{E}}_j}$, and for which each exogenous variable has at most one child (Pearl, 2009). They are interesting since they satisfy several Markov properties. Acyclic SCMs have a considerable technical advantage: they are always uniquely solvable, as we will see in the next section.

2.3 Induced measure for uniquely solvable SCMs

In this work, we will focus on SCMs that are what we call "uniquely solvable". Unique solvability can be thought of as a consistency requirement which allows us to talk about the joint probability measure entailed by an SCM.

Definition 9 *An SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is called* uniquely solvable *if for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ the structural equation*

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

has a unique solution $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.

For two equivalent SCMs, if one of the SCMs is uniquely solvable then the other is also uniquely solvable and hence unique solvability descends to the set of equivalence classes.

Proposition 10 *Every acyclic SCM is uniquely solvable.*

Proof. The proof proceeds recursively, following a topological ordering of the endogenous variables $\mathcal{I}$ in $\mathcal{M}$, where we start from the source node. As the value of an endogenous variable is (for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every value of the exogenous variables) completely determined by its direct causes, the values of all endogenous variables are $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.e. completely determined by the values of the exogenous variables. Therefore, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ the equation $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})$ has a unique solution. □

However, cyclic SCMs need not be uniquely solvable, as for a non-negligible set of values for the exogenous variables, there can be zero or multiple solutions for the endogenous variables.

Example 11 *A simple cyclic SCM is given by $\mathcal{M} = \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$, where*

$$f_1(\boldsymbol{x}, e_1) = \frac{1}{2}x_2 + e_1$$
$$f_2(\boldsymbol{x}, e_1) = x_1 \,,$$

$\mathbb{P}_{\mathbb{R}}$ is the standard-normal measure on $\mathbb{R}$, and where we used the notation $\boldsymbol{n}$ to denote $\{1, \dots, n\}$ for $n \in \mathbb{N}$. Unique solvability can be read from

[7] In the structural equation model (SEM) literature, acyclic SCMs are referred to as *recursive* SEMs and cyclic SCMs as *non-recursive* SEMs.

the structural equations, which gives for every $e \in \mathbb{R}$ the unique solution $(2e_1, 2e_1) \in \mathbb{R}^2$. Changing the causal mechanism $\boldsymbol{f}$ to

$$
\begin{aligned}
f_1(\boldsymbol{x}, e_1) &= x_2 + e_1 \\
f_2(\boldsymbol{x}, e_1) &= x_1 - e_1
\end{aligned}
$$

gives an example of a cyclic SCM that is not uniquely solvable, since the structural equations give for every $e_1 \in \mathbb{R}$ the set of solutions $\{(t, t - e_1) \in \mathbb{R}^2 : t \in \mathbb{R}\}$. Another example is obtained by changing the causal mechanism $\boldsymbol{f}$ to

$$
\begin{aligned}
f_1(\boldsymbol{x}, e_1) &= x_2 + e_1 \\
f_2(\boldsymbol{x}, e_1) &= x_1 + e_1 \,.
\end{aligned}
$$

This cyclic SCM is not uniquely solvable, since for all $e_1 \neq 0$ the structural equations have no solutions.

Unique solvability allows us to define a unique measure induced by the SCM. In the acyclic case, this is straightforward, but in the general case, some additional measure-theoretic details have to be addressed:

Lemma 12 *A uniquely solvable SCM induces a measurable mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$:*

$$\boldsymbol{g}(\boldsymbol{e}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{e}), \boldsymbol{e}).$$

This $\boldsymbol{g}$ is unique up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. Moreover, two equivalent uniquely solvable SCMs induce the same measurable mapping $\boldsymbol{g}$ up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set.

Proof. This follows from Lemma 32, where we take $\mathcal{L} = \mathcal{I}$ and $\mathcal{O} = \emptyset$. $\square$

Unique solvability and hence the existence of the mapping $\boldsymbol{g}$ depends on the distribution $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$:

Example 13 *Consider a cyclic SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ with the following causal mechanism:*

$$
\begin{aligned}
f_1(\boldsymbol{x}, e_1) &= x_2 + 1 \\
f_2(\boldsymbol{x}, e_1) &= x_1 x_2 - e_1
\end{aligned}
$$

and take for $\mathbb{P}_{\mathbb{R}}$ the standard-normal measure on $\mathbb{R}$. Then $\mathcal{M}$ is not uniquely solvable, since for every e_1 the structural equations give two solutions $(1 + \sqrt{e_1}, \sqrt{e_1})$ and $(1 - \sqrt{e_1}, -\sqrt{e_1})$. However, if we change the standard-normal measure $\mathbb{P}_{\mathbb{R}}$ to the Dirac measure δ_0, then $\mathcal{M}$ is uniquely solvable, and hence induces a measurable mapping $\boldsymbol{g} : \mathbb{R}^1 \to \mathbb{R}^2$, which can be defined by $e_1 \mapsto (1, 0)$.

Unique solvability is important because:

Corollary 14 *A uniquely solvable SCM has a solution, and all solutions of a uniquely solvable SCM have the same distribution.*

Proof. Let $\boldsymbol{g}$ be the mapping as in Lemma 12. First we show the existence of a solution: the pair $(\boldsymbol{E}^\dagger, \boldsymbol{X}^\dagger)$ with random variables $\boldsymbol{E}^\dagger : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}} : \boldsymbol{e} \mapsto \boldsymbol{e}$ and $\boldsymbol{X}^\dagger : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}} : \boldsymbol{e} \mapsto \boldsymbol{g}(\boldsymbol{e})$, with underlying probability space $(\boldsymbol{\mathcal{E}}, \mathcal{B}(\boldsymbol{\mathcal{E}}), \mathbb{P}_{\boldsymbol{\mathcal{E}}})$, is a solution. Now suppose that $(\boldsymbol{E}, \boldsymbol{X})$ is a solution of $\mathcal{M}$. Let

$$\boldsymbol{\mathcal{E}}_0 := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \neg(\exists! \boldsymbol{x} \in \boldsymbol{\mathcal{X}} : \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}))\}.$$

Then there exists an $\boldsymbol{\mathcal{N}} \in \mathcal{B}(\boldsymbol{\mathcal{E}})$ such that $\boldsymbol{\mathcal{E}}_0 \subseteq \boldsymbol{\mathcal{N}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{N}}) = 0$. Take the subspace $\boldsymbol{\mathcal{E}}^* = \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{N}}$ of $\boldsymbol{\mathcal{E}}$, then by Lemma 7.2.2 in Cohn (2013) we have:

$$\mathcal{B}(\boldsymbol{\mathcal{E}}^*) = \mathcal{B}(\boldsymbol{\mathcal{E}}) \cap \boldsymbol{\mathcal{E}}^* := \{\boldsymbol{B} \cap \boldsymbol{\mathcal{E}}^* \, : \, \boldsymbol{B} \in \mathcal{B}(\boldsymbol{\mathcal{E}})\}.$$

Note that $\boldsymbol{E} \in \boldsymbol{\mathcal{E}}^*$ a.s. and that $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$ a.s.. Therefore, $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{E})$ a.s.. Therefore, the distribution of $(\boldsymbol{E}, \boldsymbol{X})$ equals the distribution of $(\boldsymbol{E}, \boldsymbol{g}(\boldsymbol{E}))$, which is the push-forward[8] of the distribution of $\boldsymbol{E}$ under $\boldsymbol{e} \mapsto (\boldsymbol{e}, \boldsymbol{g}(\boldsymbol{e}))$, which equals the push-forward of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ under that mapping. □

For a uniquely solvable SCM $\mathcal{M}$, we get the marginal distribution over $\boldsymbol{\mathcal{X}}$ by pushing forward the probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ along the induced mapping $\boldsymbol{g}$, as defined in Lemma 12. From Corollary 14 it follows that for every solution $(\boldsymbol{E}, \boldsymbol{X})$ this marginal distribution coincides with the distribution $\mathbb{P}^{\boldsymbol{X}}$, which we call the induced *observational distribution* of $\mathcal{M}$. Below, we will always implicitly assume the existence of some solution $(\boldsymbol{E}, \boldsymbol{X})$ whenever an SCM is uniquely solvable.

2.4 Interventions

So far, we have not discussed the causal semantics of structural causal models. Following Pearl (2009), we focus here on a particular idealized class of interventions.[9]

Definition 15 *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 $\boldsymbol{\xi}_I \in \prod_{i \in I} \mathcal{X}_i$, the* perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *maps $\mathcal{M}$ to the intervened model $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ 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 5) on the set of all SCMs and hence this mapping descends to the set of equivalence classes of SCMs.

We can define an operation $\mathrm{do}(I)$ that operates on directed graphs:

[8]If (X, Σ_X) and (Y, Σ_Y) are measurable spaces, $f : X \to Y$ is a measurable function, and μ is a measure on (X, Σ_X), then one defines the *push-forward* measure of μ as the measure $\mu \circ f^{-1}$ on (Y, Σ_Y), where $f^{-1}(A) := \{x \in X : f(x) \in A\}$ denotes the pre-image under f of $A \subseteq Y$.

[9]Other types of interventions, like probabilistic interventions, mechanism changes, activity interventions, fat-hand interventions, etc. are at least as interesting, but we will not consider those here.

Definition 16 *Given a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a subset $I \subseteq \mathcal{V}$, we define the intervened graph* $\mathrm{do}(I)(\mathcal{G}) = (\mathcal{V}, \tilde{\mathcal{E}})$ *where* $\tilde{\mathcal{E}} := \mathcal{E} \setminus (\mathcal{V} \times I)$.

The two notions of intervention are compatible:

Proposition 17 *Given an SCM $\mathcal{M}$ and a subset $I \subseteq \mathcal{I}$ of endogenous variables and a value $\boldsymbol{\xi}_I \in \prod_{i \in I} \mathcal{X}_i$, then $\big(\mathcal{G}^a \circ \mathrm{do}(I, \boldsymbol{\xi}_I)\big)(\mathcal{M}) = \big(\mathrm{do}(I) \circ \mathcal{G}^a\big)(\mathcal{M})$.*

Proof. The $\mathrm{do}(I, \boldsymbol{\xi}_I)$ operation on $\mathcal{M}$ removes the functional dependence of $\boldsymbol{x}$ and $\boldsymbol{e}$ from the f_j components for $j \in I$ and hence the corresponding directed edges. □

From Corollary 14 we know that all solutions of a uniquely solvable SCM induce the same distribution. In general, the induced distribution of the solutions of the structural equations of a uniquely solvable SCM and of a uniquely solvable intervened SCM differ. For a uniquely solvable SCM $\mathcal{M}$ the induced marginal distribution over $\boldsymbol{\mathcal{X}}$ was called the observational distribution, denoted by $\mathbb{P}_{\mathcal{M}}^{\boldsymbol{X}}$.[10] Similarly, we will call the induced marginal distribution over $\boldsymbol{\mathcal{X}}$ of a uniquely solvable intervened SCM $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}$ the induced *interventional distribution* of $\mathcal{M}$ under the perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$, and is denoted by $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}}^{\boldsymbol{X}}$.[11] Note doing a perfect intervention on an SCM does not require the SCM to be uniquely solvable. When it is uniquely solvable, it may not be uniquely solvable anymore after a perfect intervention (see Example 19). We have the following elementary properties:

Lemma 18 *For any SCM:*

1. *Perfect interventions on disjoint subsets of endogenous variables commute.*
2. *Acyclicity is preserved under perfect intervention.*

Proof. The first statement follows directly from the definitions. For the second statement, note that a perfect intervention can only remove causal relations, and therefore will never introduce a cycle. □

As acyclic SCMs are uniquely solvable it follows that acyclic SCMs are always uniquely solvable under any perfect intervention. However, unique solvability is not preserved in general: a cyclic uniquely solvable SCM can become not uniquely solvable by a perfect intervention, as the next example shows.

Example 19 *Consider a cyclic uniquely solvable SCM $\mathcal{M} = \langle \mathbf{3}, \mathbf{3}, \mathbb{R}^3, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$ with the following causal mechanism:*

$$
\begin{aligned}
f_1(\boldsymbol{x}, \boldsymbol{e}) &= x_2 + e_1 \\
f_2(\boldsymbol{x}, \boldsymbol{e}) &= x_1 + x_3 + e_2 \\
f_3(\boldsymbol{x}, \boldsymbol{e}) &= -x_1 + e_3
\end{aligned}
\tag{2}
$$

[10] If it is clear from which SCM the solutions are, we can neglect the subscript $\mathcal{M}$ in the notation.

[11] 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 probability distributions $\mathbb{P}_{\mathcal{M}}^{\boldsymbol{X}}$ and $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}}^{\boldsymbol{X}}$ respectively, whenever they exist.

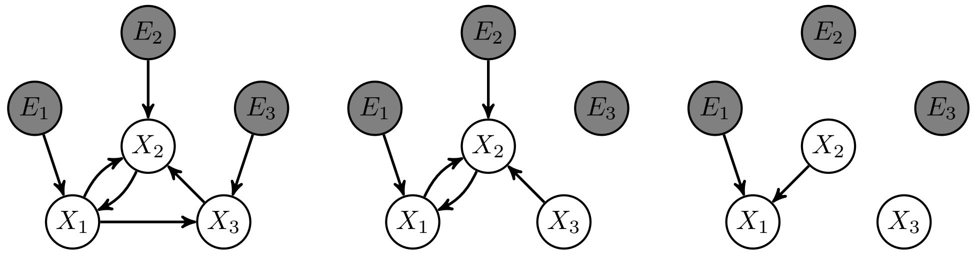

Figure 1: The augmented graph $\mathcal{G}^a(\mathcal{M})$ of the cyclic SCM (left), after first intervention $\mathrm{do}(\{3\}, 1)$ (middle), and after second intervention $\mathrm{do}(\{2\}, 1)$ (right).

and take for $\mathbb{P}_{\mathbb{R}^3}$ the standard-normal measure on $\mathbb{R}^3$. The augmented graph $\mathcal{G}^a(\mathcal{M})$ is depicted in Figure 1.[12] *The perfect intervention $\mathrm{do}(\{3\}, 1)$ gives the intervened causal mechanism*

$$\begin{aligned} \tilde{f}_1(\boldsymbol{x}, \boldsymbol{e}) &= x_2 + e_1 \\ \tilde{f}_2(\boldsymbol{x}, \boldsymbol{e}) &= x_1 + x_3 + e_2 \\ \tilde{f}_3(\boldsymbol{x}, \boldsymbol{e}) &= 1 \end{aligned} \tag{3}$$

which is not uniquely solvable anymore. Hence a perfect intervention can destroy the unique solvability property. The reverse is also possible: doing another perfect intervention $\mathrm{do}(\{2\}, 1)$ on $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ gives again a uniquely solvable SCM. Note that the preservation of unique solvability under perfect interventions depends on the distribution over the exogenous variables. If one takes the product of Dirac measures δ_0 and δ_{-1}, for resp. e_1 and e_2, then the above example will be uniquely solvable.

We now prove a very intuitive notion: only an intervention on an ancestor of a variable can change the distribution of that variable. In the cyclic case, there is a subtlety as can be seen in Example 21: in order for this to hold, an extra assumption has to be made. First note that from the definition of ancestors it follows that one can always choose an equivalent representation of $\boldsymbol{f}$ such that $\boldsymbol{f}_{\mathrm{an}(i)} : \boldsymbol{\mathcal{X}}_{\mathrm{an}(i)} \times \boldsymbol{\mathcal{E}}_{\mathrm{an}(i)} \to \boldsymbol{\mathcal{X}}_{\mathrm{an}(i)}$ does not depend on non-ancestors of i.[13]

Proposition 20 *Let $\mathcal{M}$ be an SCM and let $i \in \mathcal{I}$. Assume that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ the structural equation*

$$\boldsymbol{x}_{\mathrm{an}(i)} = \boldsymbol{f}_{\mathrm{an}(i)}(\boldsymbol{x}_{\mathrm{an}(i)}, \boldsymbol{e}_{\mathrm{an}(i)})$$

has a unique solution $\boldsymbol{x}_{\mathrm{an}(i)}$, then an intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I} \setminus \mathrm{an}(i)$ will not change the distribution on $\boldsymbol{\mathcal{X}}_{\mathrm{an}(i)}$ (assuming that both $\mathcal{M}$ and $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ are uniquely solvable).

[12] For the graphical representation of a graph, we will stick to the common convention to use random variable, with the index set as subscript, instead of using the index set itself. This does not imply that these random variables will always exist as a solution of the SCM, since the SCM may not have any solution.

[13] Whenever it is clear from the context we use the notation $\mathrm{an}(i) \subseteq \mathcal{I} \cup \mathcal{J}$ for both $\mathrm{an}(i) \cap \mathcal{I}$ and $\mathrm{an}(i) \cap \mathcal{J}$. For example, we will write $\boldsymbol{\mathcal{X}}_{\mathrm{an}(i)}$ and $\boldsymbol{\mathcal{E}}_{\mathrm{an}(i)}$ instead of $\boldsymbol{\mathcal{X}}_{\mathrm{an}(i)\cap\mathcal{I}}$ and $\boldsymbol{\mathcal{E}}_{\mathrm{an}(i)\cap\mathcal{J}}$ respectively.

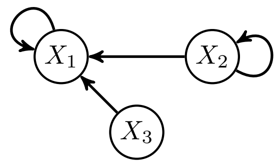

Figure 2: Cyclic SCM that violates the intuitive notion (see Example 21).

Proof. According to the assumption, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ the equation

$$\boldsymbol{x}_{\mathrm{an}(i)} = \boldsymbol{f}_{\mathrm{an}(i)}(\boldsymbol{x}_{\mathrm{an}(i)}, \boldsymbol{e}_{\mathrm{an}(i)})$$

has a unique solution $\boldsymbol{x}_{\mathrm{an}(i)}$. Note that an intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ with $I \subseteq \mathcal{I} \setminus \mathrm{an}(i)$ will keep the components $\boldsymbol{f}_{\mathrm{an}(i)}$ invariant. Hence, when both $\mathcal{M}$ and $\mathcal{M}' = \mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ are uniquely solvable, then $(\boldsymbol{g}^{\mathcal{M}})_{\mathrm{an}(i)} = (\boldsymbol{g}^{\mathcal{M}'})_{\mathrm{an}(i)}$ up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. Therefore, the interventional distribution on $\boldsymbol{\mathcal{X}}_{\mathrm{an}(i)}$ according to $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ equals the observational distribution on $\boldsymbol{\mathcal{X}}_{\mathrm{an}(i)}$ according to $\mathcal{M}$. $\square$

The assumption made in Proposition 20 is equivalent to what we later will call uniquely solvable with respect to $\mathrm{an}(i)$. In the following example (inspired by an example in Neal (2000)) we see what can go wrong if the assumption is not satisfied.

Example 21 *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, 1, \boldsymbol{f}, \mathbb{P}_1 \rangle$ *with the causal mechanism:*

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

where $\mathbf{1}_{\{0\}}$ *denotes the indicator function and* $\mathbb{P}_1$ *is the trivial probability measure over the point, that is it satisfies* $\mathbb{P}_1(\{1\}) = 1$. *Note* $\mathcal{M}$ *and* $\mathcal{M}_{\mathrm{do}(3,c)}$ *for* $c \in \mathbb{R}$ *are uniquely solvable. One can check that a perfect intervention on* x_3 *changes the distribution of* x_2 *in general, even though* x_3 *is not an ancestor of* x_2*, as can be seen from the directed graph in Figure 2. This strange behavior can only happen if an SCM does not satisfy the extra assumption. Note that* $\{x_2\} = \mathrm{an}(x_2)$*, but the equation* $x_2 = x_2$ *has no unique solution.*

2.5 Equivalence relations between SCMs

In the causality literature, exogenous variables are typically taken to be latent, even though one could also treat them as observed variables or as partially observed. The following terminology reflects that assumption in that it ignores the distribution of the exogenous variables and only considers the marginal distribution of the endogenous variables.

We consider two SCMs to be *observationally equivalent* if they cannot be distinguished based on their induced observational distributions:

Definition 22 *Two SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are* observationally equivalent with respect to $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$ *if $\mathcal{X}_\mathcal{O} = \tilde{\mathcal{X}}_\mathcal{O}$ and they are either (i) both uniquely solvable and they induce the same marginal observational distribution $\mathbb{P}^{\boldsymbol{X}_\mathcal{O}}$, or (ii) both not uniquely solvable. They are called* observationally equivalent *if they are observationally equivalent with respect to $\mathcal{I} = \tilde{\mathcal{I}}$, i.e., with respect to all endogenous variables.*

We consider two SCMs to be *interventionally equivalent* if they induce the same distributions under any perfect intervention:

Definition 23 *Two SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are* interventionally equivalent with respect to $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$ *if for any $I \subseteq \mathcal{O}$ and any 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 $\mathcal{O}$. They are called* interventionally equivalent *if they are interventionally equivalent with respect to $\mathcal{I} = \tilde{\mathcal{I}}$.*

Interventionally equivalent SCMs induce the same interventional distributions. However in contrast to equivalent SCMs, interventionally equivalent SCMs do in general not have the same (augmented) graph, as we will see later, e.g. in the example given by Figure 5. Considering empty perfect interventions ($I = \emptyset$) as a special case of a perfect intervention, interventional equivalence implies observational equivalence. Interventional equivalence is a strictly stronger notion than observational equivalence, as the following well-known example illustrates.

Example 24 *Consider the uniquely solvable SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{2}, \mathbb{R}^2, \mathbb{R}^2, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^2} \rangle$ with causal mechanism*

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

where $\alpha \in \mathbb{R}$ and with independent exogenous variables E_1, E_2 with normal distributions

$$E_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$E_2 \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

The augmented graph is depicted in Figure 3. Consider also the uniquely solvable SCM $\tilde{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \mathbb{R}^2, \mathbb{R}^2, \tilde{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^2} \rangle$ with causal mechanism

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = \beta x_2 + \tilde{e}_1$$
$$f_2(\boldsymbol{x}, \boldsymbol{e}) = \tilde{e}_2,$$

where

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

and with independent exogenous variables $\tilde{E}_1, \tilde{E}_2$ with normal distributions

$$\tilde{E}_1 \sim \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2)$$
$$\tilde{E}_2 \sim \mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2),$$

where

$$\tilde{\mu}_2 = \alpha \mu_1 + \mu_2, \quad \tilde{\sigma}_2^2 = \alpha^2 \sigma_1^2 + \sigma_2^2,$$
$$\tilde{\mu}_1 = (1 - \alpha\beta)\mu_1 - \beta\mu_2, \quad \tilde{\sigma}_1^2 = (1 - \alpha\beta)^2 \sigma_1^2 + \beta^2 \sigma_2^2.$$

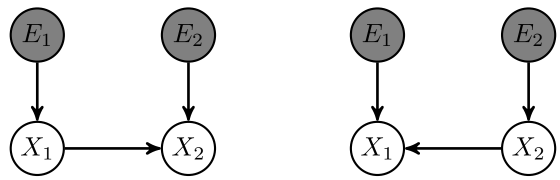

Figure 3: The augmented graph of the SCM $\mathcal{M}$ (left) and $\tilde{\mathcal{M}}$ (right). They are observational, but not interventional equivalent.

Then $\mathcal{M}'$ and $\mathcal{M}$ are observationally equivalent, but not interventionally equivalent.

The definition of observational and interventional equivalence both define an equivalence relation on the set of all SCMs. Note that two SCMs are observational and interventional equivalent if they are with respect to the set of endogenous variables, meaning that between those SCMs the index set of exogenous variables, the space of exogenous variables, the exogenous probability distribution and the causal mechanism may differ. The set of observational and interventional equivalence classes of SCMs are denoted by $\mathbf{SCM}_{obs}$ and $\mathbf{SCM}_{int}$ respectively. These set of equivalence classes are related in the following way:

$$\begin{aligned} & \mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are equivalent} \\ \Longrightarrow\ & \mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are interventionally equivalent} \\ \Longrightarrow\ & \mathcal{M} \text{ and } \tilde{\mathcal{M}} \text{ are observationally equivalent.} \end{aligned}$$

Altogether this gives the following partial order on the set of different equivalence classes $\mathbf{SCM} \leq \mathbf{SCM}_{eq} \leq \mathbf{SCM}_{int} \leq \mathbf{SCM}_{obs}$, where '$\leq$' means finer-than (and the equivalence relation on $\mathbf{SCM}$ is the trivial one). For a subset $\mathcal{O}$ we can similarly define $\mathbf{SCM} \leq \mathbf{SCM}_{int}(\mathcal{O}) \leq \mathbf{SCM}_{obs}(\mathcal{O})$, where $\mathbf{SCM}_{obs}(\mathcal{O})$ and $\mathbf{SCM}_{int}(\mathcal{O})$ mean respectively the set of observational and interventional equivalence classes w.r.t. $\mathcal{O}$.

3 Marginalization

In this section we define a marginalization operation on SCMs that marginalizes over a subset $\mathcal{L} \subset \mathcal{I}$ of endogenous variables. Intuitively, the idea is that we would like to treat the subsystem $\mathcal{L}$ as a "black box", and only describe how the rest of the system interacts with it. Thereby we completely remove the representation of the internals of the subsystem, preserving only the essential input-output characteristics of it. This is somewhat analogous to marginalizing a probability distribution on a set of variables down to a distribution on a subset of those variables, hence the name "marginalization". We will show that if the part of the causal mechanism that describes the causal relations in the subsystem $\mathcal{L}$ (i.e., the restriction of the causal mechanism to $\mathcal{L}$) satisfies a certain unique solvability condition (intuitively: it gives a unique output for any possible input), we can effectively remove this subsystem of endogenous variables from the model, treating it as a black box. An important property of this marginalization operation is that it preserves the causal semantics, meaning that the interventional distributions induced by the SCM are identical to those induced by its marginalization.

We will start by considering the more intuitive notion of marginalizing over endogenous variables for acyclic SCMs, and then consider the more abstract definition that works for arbitrary SCMs.

For an acyclic SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ we can always marginalize over an endogenous variable $\ell \in \mathcal{I}$ by first choosing a representation $\tilde{\boldsymbol{f}}$ such that $\tilde{f}_\ell$ does not depend on x_ℓ, and then defining the marginal causal mechanism $\boldsymbol{f}_{\mathrm{marg}(\ell)}$ of $\mathcal{M}_{\mathrm{marg}(\ell)} = \langle \mathcal{I} \setminus \{\ell\}, \mathcal{J}, \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\{\ell\}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}_{\mathrm{marg}(\ell)}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ as

$$\boldsymbol{f}_{\mathrm{marg}(\ell)}(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, \boldsymbol{e}) := \tilde{\boldsymbol{f}}_{\mathcal{I}\setminus\{\ell\}}(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, \tilde{f}_\ell(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, \boldsymbol{e}), \boldsymbol{e}).$$

In other words, we simply substitute the causal mechanism of ℓ into the other causal mechanisms. Note that the marginal causal mechanism on $\boldsymbol{x}_{\setminus \ell}$ no longer depends on x_ℓ. In the graphical representation, this marginalization operation replaces every edge $\ell \to j$ for $j \in \mathcal{I} \setminus \{\ell\}$ by the set of directed edges $i \to j$ for every $i \in \mathcal{I} \setminus \{\ell\} \cup \mathcal{J}$ such that there is a directed path $i \to \ell \to j$, second it removes all the remaining edges pointing in or out ℓ and finally it removes ℓ itself. From this we can see that marginalizing over an endogenous variable does not introduce any cycle, and hence:

Lemma 25 *Marginalizing over an endogenous variable $\ell \in \mathcal{I}$ preserves acyclicity.*

Example 26 *Consider the acyclic SCM $\mathcal{M} = \langle \mathbf{4}, \mathbf{1}, \mathbb{R}^4, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_\mathbb{R} \rangle$ with the following causal mechanism:*

$$\begin{aligned} f_1(\boldsymbol{x}, e_1) &= 1 \\ f_2(\boldsymbol{x}, e_1) &= x_1 + e_1 \\ f_3(\boldsymbol{x}, e_1) &= x_2 \\ f_4(\boldsymbol{x}, e_1) &= x_2 \end{aligned}$$

and where $\mathbb{P}_\mathbb{R}$ is the standard-normal measure on $\mathbb{R}$. Its augmented graph is depicted in Figure 4. Marginalizing over x_2 yields by substitution the marginal causal mechanism:

$$\begin{aligned} f_{\mathrm{marg}(2),1}(\boldsymbol{x}, e_1) &= 1 \\ f_{\mathrm{marg}(2),3}(\boldsymbol{x}, e_1) &= x_1 + e_1 \\ f_{\mathrm{marg}(2),4}(\boldsymbol{x}, e_1) &= x_1 + e_1. \end{aligned}$$

Marginalizing over x_1 now gives the marginal causal mechanism:

$$\begin{aligned} (f_{\mathrm{marg}(2)})_{\mathrm{marg}(1),3}(\boldsymbol{x}, e_1) &= 1 + e_1 \\ (f_{\mathrm{marg}(2)})_{\mathrm{marg}(1),4}(\boldsymbol{x}, e_1) &= 1 + e_1 \end{aligned}$$

Note that first marginalizing x_1 and then x_2 would yield the same marginal causal mechanism.

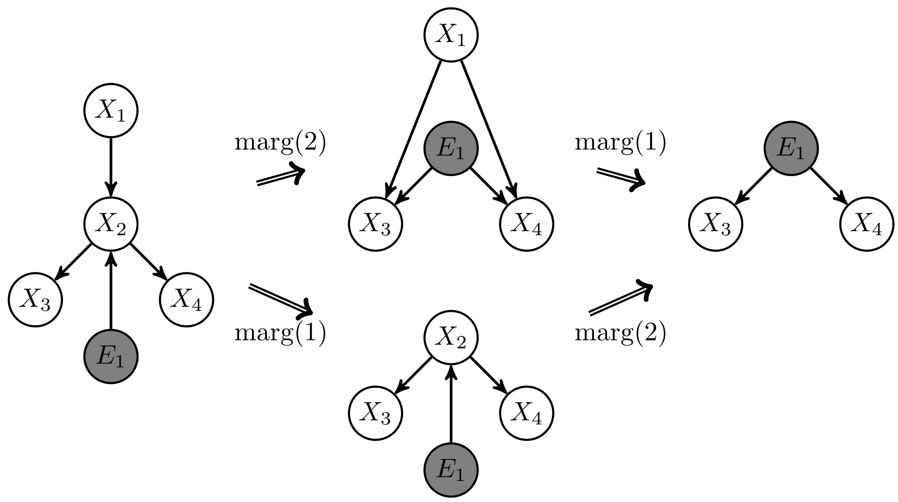

Figure 4: Example of commutatitivity property of marginalization (see Example 26).

We will prove in Lemma 38 that we can uniquely marginalize over any subset of variables for acyclic SCMs. However, in the cyclic case we cannot in general remove the functional dependence of an endogenous variable by substitution, as can be seen by the following example:

Example 27 *Consider the cyclic SCM* $\mathcal{M} = \langle \mathbf{2}, \mathbf{2}, \mathbb{R}^2, \mathbb{R}^2, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^2} \rangle$ *with the following causal mechanism:*

$$f_1(\boldsymbol{x}, \boldsymbol{e}) = x_2 + e_1$$
$$f_2(\boldsymbol{x}, \boldsymbol{e}) = \frac{1}{2} x_2 + e_2 \,.$$

where we take for $\mathbb{P}_{\mathbb{R}^2}$ *the standard multivariate normal measure on* $\mathbb{R}^2$*. Because of the self-loop on* x_2*, there exists no equivalent representation* $\tilde{\boldsymbol{f}}$ *such that* $\tilde{f}_\ell$ *does not depend on* x_2*, that is there is a self-loop on* x_2*. However the structural equation* $x_2 = \frac{1}{2} x_2 + e_2$ *gives a unique solution for* x_2 *given* e_2*. Substituting this into the structural equation of* x_1 *yields* $x_1 = 2e_2 + e_1$*. As we will see below*

$$f_{\mathrm{marg}(2),1} := 2e_2 + e_1$$

is indeed the marginal causal mechanism for x_1*. Uniquely solving first the structural equation for* x_2*, that is the one with the self-loop, allowed us to define the marginal causal mechanism.*

One may guess that for a cyclic SCM, one can simply first uniquely solve all the structural equations that include self-loops and then perform the substitutions as in the acyclic case. However, this does not work, as can be seen from the following example.

Example 28 *Consider the cyclic SCM $\mathcal{M} = \langle \mathbf{3}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ with the following causal mechanism:*

$$\begin{aligned} f_1(\boldsymbol{x}, e_1) &= 2x_1 - x_2 - e_1 \\ f_2(\boldsymbol{x}, e_1) &= x_1 \\ f_3(\boldsymbol{x}, e_1) &= x_2 \,. \end{aligned}$$

where we take $\mathbb{P}_{\mathbb{R}}$ the standard-normal measure on $\mathbb{R}$. Suppose we want to marginalize over x_1 and x_2, then we can start for example by marginalizing over x_1. Since this SCM contains a self-loop at x_1 we first need to uniquely solve the structural equation for x_1, before we can perform the substitution. This gives us a marginal SCM with causal mechanism:

$$\begin{aligned} f_{\mathrm{marg}(1),2}(\boldsymbol{x}, e_1) &= x_2 + e_1 \\ f_{\mathrm{marg}(1),3}(\boldsymbol{x}, e_1) &= x_2 \,. \end{aligned}$$

Marginalizing in this way over x_1 introduces a new self-loop. Now trying to marginalize over x_2, by uniquely solving the structural equation with self-loop, does not work, since the structural equation $x_2 = x_2 + e_1$ does not have any solution for a given e_1, hence it seems that marginalizing over x_1 and x_2 is not possible.

Acyclic SCMs have the nice property that for every $\ell \in \mathcal{I}$ for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}} \in \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\{\ell\}}$ the structural equation

$$x_\ell = f_\ell(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, x_\ell, \boldsymbol{e})$$

has a unique solution $x_\ell \in \mathcal{X}_\ell$, since we can always find a representation of f_ℓ that does not depend on x_ℓ, that is $\tilde{\boldsymbol{f}}(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, \boldsymbol{e})$, as $\mathcal{M}$ has no self-loops by definition. Intuitively, the structural equation $x_\ell = f_\ell(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, x_\ell, \boldsymbol{e})$ defines a "subsystem" in the sense that for a given input $(\boldsymbol{x}_{\mathcal{I}\setminus\{\ell\}}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\{\ell\}} \times \boldsymbol{\mathcal{E}}$ we get a unique output $x_\ell \in \mathcal{X}_\ell$.

In general, for an SCM and a given subset $\mathcal{L} \subseteq \mathcal{I}$ of endogenous variables and its complement $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$, the structural equations $\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{L}}, \boldsymbol{e})$ define a "subsystem", for which an example is depicted by the blackbox in Figure 5. We consider $(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}$ as the input of the subsystem, and $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ (satisfying $\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{L}}, \boldsymbol{e})$) as the output of the subsystem. As in the acyclic case, for each input we would like the subsystem to give a unique output. If this is the case, then by marginalizing we can effectively remove this subystem of endogenous variables from the model, which will lead to a marginal SCM that is interventionally equivalent to the original one, which we will prove in Theorem 43.

Definition 29 *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 $\mathcal{L} \subseteq \mathcal{I}$ (and its complement $\mathcal{O} := \mathcal{I} \setminus \mathcal{L}$) of endogenous variables, we call the SCM $\mathcal{M}$* uniquely solvable with respect to $\mathcal{L}$ *if for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{\xi}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ the structural equation*

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

has a unique solution $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}$.

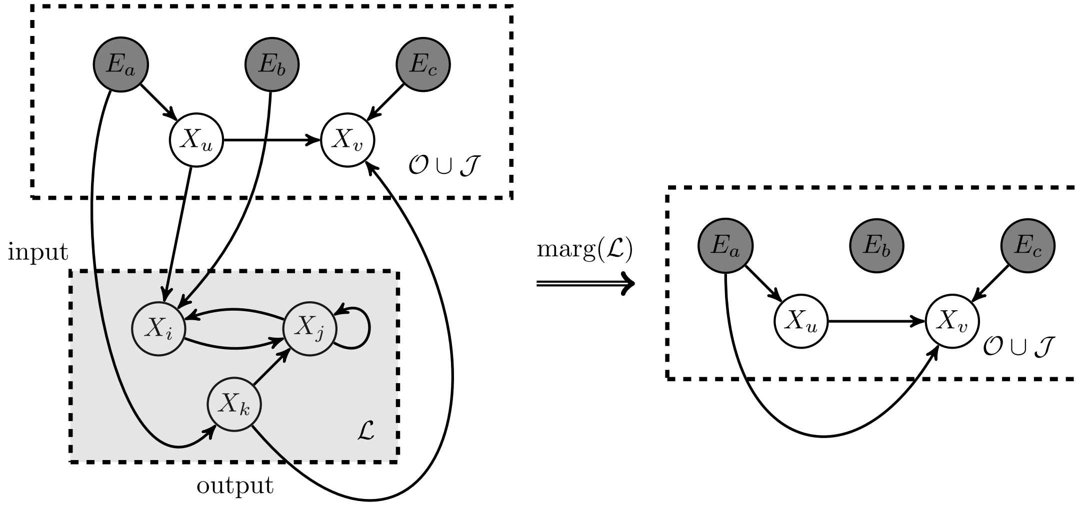

Figure 5: Example of a subsystem $\mathcal{L}$ of an SCM on the left, depicted as a blackbox, which for a given input $(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}$ has a unique output $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}$. Marginalizing over this subsystem gives the marginal SCM on the right.

For two equivalent SCMs, whenever one of them is uniquely solvable w.r.t. $\mathcal{L}$, the other is too. The case where an SCM is uniquely solvable w.r.t. $\mathcal{I}$ itself means exactly that the SCM is uniquely solvable. Note that solvability does not in general imply unique solvability w.r.t. $\mathcal{L}$, acyclicity however does imply unique solvability w.r.t. $\mathcal{L}$ for any $\mathcal{L} \subseteq \mathcal{I}$.

The following example illustrates that unique solvability is in general not preserved under union and intersection.

Example 30 *Consider the SCM $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, 1, \boldsymbol{f}, \mathbb{P}_1 \rangle$ with the causal mechanism:*

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

where $\mathbf{1}_{\{0\}}$ denotes the indicator function and $\mathbb{P}_1$ is the trivial probability measure over the point. Take $\mathcal{L}_1 = \{1, 2\}$ and $\mathcal{L}_2 = \{2, 3\}$, then $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1$ and $\mathcal{L}_2$, however $\mathcal{M}$ is neither uniquely solvable w.r.t. $\mathcal{L}_1 \cup \mathcal{L}_2$ nor $\mathcal{L}_1 \cap \mathcal{L}_2$. Moreover, consider the disjoint subsets $\{1\}$ and $\{2\}$ of $\mathbf{3}$, then $\mathcal{M}$ is uniquely solvable w.r.t. the union $\mathcal{L}_1$, however it is not uniquely solvable w.r.t. $\{1\}$ and $\{2\}$.

However we have the following property:

Proposition 31 *Given an SCM $\mathcal{M}$ and disjoint subsets $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{I}$. If $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1$ and $\mathcal{L}_2$, then it is uniquely solvable w.r.t. their union $\mathcal{L}_1 \cup \mathcal{L}_2$.*

Proof. This follows straightforward from the definition. $\square$

A uniquely solvable SCM induces a measurable mapping $\boldsymbol{g}$, and similarly we have:

Lemma 32 *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 $\mathcal{L} \subseteq \mathcal{I}$. If $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then this induces a measurable mapping $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ we have:*

$$\forall \boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} : \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{e}).$$

This mapping $\boldsymbol{g}_{\mathcal{L}}$ is unique up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. Moreover, two equivalent SCMs that are uniquely solvable w.r.t. $\mathcal{L}$ induce the same measurable mapping $\boldsymbol{g}_{\mathcal{L}}$ up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set.

Proof. The non-trivial part of the proof is to show that $\boldsymbol{g}_{\mathcal{L}}$ is measurable. Consider an SCM that is uniquely solvable w.r.t. $\mathcal{L}$ and assume without loss of generality that $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{E}}$ are Polish spaces with Borel σ-algebras $\mathcal{B}(\boldsymbol{\mathcal{X}})$ and $\mathcal{B}(\boldsymbol{\mathcal{E}})$. Let

$$\boldsymbol{\mathcal{E}}_0 := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \neg(\forall \boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \exists! \boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}} : \boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{L}}, \boldsymbol{e}))\}.$$

Then there exists an $\boldsymbol{\mathcal{N}} \in \mathcal{B}(\boldsymbol{\mathcal{E}})$ such that $\boldsymbol{\mathcal{E}}_0 \subseteq \boldsymbol{\mathcal{N}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{N}}) = 0$. Take the subspace $\boldsymbol{\mathcal{E}}^* = \boldsymbol{\mathcal{E}} \setminus \boldsymbol{\mathcal{N}}$ of $\boldsymbol{\mathcal{E}}$, then by Lemma 7.2.2 in Cohn (2013) we have:

$$\mathcal{B}(\boldsymbol{\mathcal{E}}^*) = \mathcal{B}(\boldsymbol{\mathcal{E}}) \cap \boldsymbol{\mathcal{E}}^* := \{\boldsymbol{B} \cap \boldsymbol{\mathcal{E}}^* \ : \ \boldsymbol{B} \in \mathcal{B}(\boldsymbol{\mathcal{E}})\}.$$

Define $\boldsymbol{h} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^* \times \boldsymbol{\mathcal{X}}_{\mathcal{L}} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}} \times \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ by

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

The function $\boldsymbol{h}$ is measurable, which follows from the measurability of $\boldsymbol{f}$.

Note that $\Delta := \{(\boldsymbol{x}_{\mathcal{L}}, \boldsymbol{x}_{\mathcal{L}}) : \boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}\}$ is a Borel subset in $\boldsymbol{\mathcal{X}}_{\mathcal{L}} \times \boldsymbol{\mathcal{X}}_{\mathcal{L}}$, since $\boldsymbol{\mathcal{X}}_{\mathcal{L}}$ is Hausdorff. Hence this implies that

$$\boldsymbol{h}^{-1}(\Delta) = \{(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}, \boldsymbol{x}_{\mathcal{L}}) \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^* \times \boldsymbol{\mathcal{X}}_{\mathcal{L}} \ : \ \boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{L}}, \boldsymbol{e})\}$$

is measurable. Moreover, since $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, we have that the equation $\boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{L}}, \boldsymbol{e})$ maps every $(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^*$ to a unique solution $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}}$, and hence this gives a unique map $\boldsymbol{g}^*_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^* \to \boldsymbol{\mathcal{X}}_{\mathcal{L}}$. The graph of $\boldsymbol{g}^*_{\mathcal{L}}$ is given by

$$\begin{aligned}
\boldsymbol{\mathcal{G}} &:= \{(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}, \boldsymbol{g}^*_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e})) \ : \ (\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^*\} \\
&= \{(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}, \boldsymbol{x}_{\mathcal{L}}) \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^* \times \boldsymbol{\mathcal{X}}_{\mathcal{L}} \ : \ \boldsymbol{x}_{\mathcal{L}} = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{x}_{\mathcal{L}}, \boldsymbol{e})\} \\
&= \boldsymbol{h}^{-1}(\Delta).
\end{aligned}$$

Since $\boldsymbol{\mathcal{G}}$ is a Borel set of $\boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}}^* \times \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ it follows directly from Proposition 8.3.4 in Cohn (2013) that the mapping $\boldsymbol{g}^*_{\mathcal{L}}$ is measurable.

Now define $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ by

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

where for $\boldsymbol{x}_0$ we can take an arbitrary point in $\boldsymbol{\mathcal{X}}_{\mathcal{L}}$. This mapping $\boldsymbol{g}_{\mathcal{L}}$ is measurable, because $\boldsymbol{g}^*_{\mathcal{L}}$ is measurable, and it is unique up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. Moreover, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ it satisfies $\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{e})$, and hence for an equivalent SCM $\tilde{\mathcal{M}}$ it also satisfies for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ that $\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{e})$. □

With this proposition at hand we can define a marginal SCM as follows:

Definition 33 *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 $\mathcal{L} \subset \mathcal{I}$ such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, a* marginalization marg($\mathcal{L}$) of the SCM $\mathcal{M}$ with respect to $\mathcal{L}$ *is an 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 "marginal" causal mechanism $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ defined by*

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

where $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}}$ is a mapping defined for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ (see Lemma 32) by

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

This definition naturally extends the concept of marginalization from the acyclic to the cyclic case. It gives both the marginal causal mechanisms of Example 26, 27, and Example 28 illustrates that marginalization may not be defined when the unique solvability condition is violated. Moreover:

Proposition 34 *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L} \subset \mathcal{I}$ such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, all marginalizations of $\mathcal{M}$ w.r.t. $\mathcal{L}$ are equivalent.*

Proof. Follows directly from Lemma 32. □

Hence for a specific $\mathcal{L} \subset \mathcal{I}$ all marginalizations marg($\mathcal{L}$) map $\mathcal{M}$ to a representative of the same equivalence class of SCMs. Moreover, marginalizing two equivalent SCMs over $\mathcal{L}$ yields two equivalent marginal SCMs. In this sense, the relation marg($\mathcal{L}$) between SCMs induces a mapping between equivalences classes of SCMs.

In the graphical representation the marginalization operation is defined as follows:

Definition 35 *Given a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a subset of nodes $\mathcal{L} \subseteq \mathcal{V}$, we define the* latent projection of $\mathcal{G}$ over $\mathcal{L}$ *as the graph* $\mathrm{marg}(\mathcal{L})(\mathcal{G}) := (\mathcal{V} \setminus \mathcal{L}, \tilde{\mathcal{E}})$ *where*

$$\tilde{\mathcal{E}} := \{i \to j : i, j \in \mathcal{V} \setminus \mathcal{L}, i \to \ell_1 \to \cdots \to \ell_n \to j \in \mathcal{G} \text{ for } n \geq 0,\ \ell_1, \ldots, \ell_n \in \mathcal{L}\}$$

The name "latent projection" is inspired from a similar construction on mixed graphs in Verma (1993). However, the latent projection defined there does not provide a mapping between SCMs, but only a mapping between mixed graphs that is shown to preserve conditional independence properties (see also Tian, 2002). We already saw examples of the latent projection: in Figure 4 for the acyclic case and in Figure 5 for the cyclic case.

Proposition 36 *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L} \subset \mathcal{I}$ such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then $\big(\mathcal{G}^a \circ \mathrm{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\mathrm{marg}(\mathcal{L}) \circ \mathcal{G}^a\big)(\mathcal{M})$.*

Proof. First, consider a representation of $\boldsymbol{f}$ such that each f_i only depends on $\mathrm{pa}(i)$. Then, for each $\ell \in \mathcal{L}$, perform the following substitution procedure: for every $k \in \mathcal{L}$ that f_ℓ depends on, substitute x_k by f_k, and let

$\tilde{f}_\ell$ be the result of this substitution procedure. Repeat this a finite number of times until each $\tilde{f}_\ell$ depends only on $\boldsymbol{x}_\mathcal{L}$ and on $\mathrm{an}(\ell) \cap (\mathrm{pa}(\mathcal{L}) \setminus \mathcal{L})$. Now let $\tilde{\boldsymbol{g}}_\mathcal{L} : \boldsymbol{\mathcal{X}}_\mathcal{O} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_\mathcal{L}$ be defined for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ by

$$\tilde{\boldsymbol{g}}_\mathcal{L}(\boldsymbol{x}_\mathcal{O}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_\mathcal{L}(\boldsymbol{x}_\mathcal{O}, \tilde{\boldsymbol{g}}_\mathcal{L}(\boldsymbol{x}_\mathcal{O}, \boldsymbol{e}), \boldsymbol{e}) \,.$$

From Lemma 32 it follows that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$

$$\forall \boldsymbol{x}_O \in \boldsymbol{\mathcal{X}}_O : \qquad \tilde{\boldsymbol{g}}_\mathcal{L}(\boldsymbol{x}_\mathcal{O}, \boldsymbol{e}) = \boldsymbol{g}_\mathcal{L}(\boldsymbol{x}_\mathcal{O}, \boldsymbol{e}).$$

Now note that the defining property of $\tilde{\boldsymbol{g}}$ implies that there exists a representation $\hat{\boldsymbol{g}}$ of $\tilde{\boldsymbol{g}}$ such that for each $\ell \in \mathcal{L}$, $\hat{g}_\ell$ only depends on $\mathrm{an}(\ell) \cap (\mathrm{pa}(\mathcal{L}) \setminus \mathcal{L})$. Using this $\hat{\boldsymbol{g}}$ to construct the marginalization $\mathrm{marg}(\mathcal{L})(\boldsymbol{M})$, we conclude that there exists a representation of $\boldsymbol{f}^{\mathrm{marg}(\mathcal{L})(\mathcal{M})}$ such that every component $(\boldsymbol{f}^{\mathrm{marg}(\mathcal{L})(\mathcal{M})})_j$ for $j \in \mathcal{O}$ depends on no other variables than $\mathrm{pa}(j) \setminus \mathcal{L}$ and

$$\bigcup_{\ell \in \mathrm{pa}(j) \cap \mathcal{L}} \mathrm{an}(\ell) \cap (\mathrm{pa}(\mathcal{L}) \setminus \mathcal{L}).$$

Therefore, the augmented graph $\mathcal{G}^a\big(\mathrm{marg}(\mathcal{L})(\mathcal{M})\big)$ is a subgraph of the latent projection $\mathrm{marg}(\mathcal{L})\big(\mathcal{G}^a(\mathcal{M})\big)$. □

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

Example 37 *Consider the SCM given by* $\mathcal{M} = \langle \mathbf{3}, \mathbf{3}, \mathbb{R}^3, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$, *where*

$$\begin{aligned} f_1(\boldsymbol{x}, \boldsymbol{e}) &= x_2 - x_3 + e_1 \\ f_2(\boldsymbol{x}, \boldsymbol{e}) &= x_1 + e_2 \\ f_3(\boldsymbol{x}, \boldsymbol{e}) &= x_1 + e_3 \end{aligned}$$

Then when we marginalize over $\{x_2, x_3\}$, *we obtain*

$$f_{\mathrm{marg}(\{x_2, x_3\})}(\boldsymbol{x}, \boldsymbol{e}) = e_2 - e_3 + e_1$$

which does not depend on x_1, *as one would expect from the latent projection.*

From the definition of a marginalization we see that marginalizing over a set of variables is only possible if we impose a unique solvability condition on the SCM. If we perform several marginalization operations after each other, then we need to impose a new unique solvability condition on the resulting SCM each time we marginalize over a set of variables. However, they can be related to each other as follows:

Lemma 38 *Given an SCM* $\mathcal{M}$ *and disjoint subsets* $\mathcal{L}_1, \mathcal{L}_2 \subset \mathcal{I}$. *Assume* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{L}_1$, *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* $\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)$.

Proof. From Lemma 32 and the unique solvability of $\mathcal{M}$ w.r.t. $\mathcal{L}_1$ it follows that there exist a unique mapping $\boldsymbol{g}_{\mathcal{L}_1} : \boldsymbol{\mathcal{X}}_{\mathcal{L}_2} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_1}$ (where $\mathcal{O} := \mathcal{I} \setminus (\mathcal{L}_1 \cup \mathcal{L}_2)$) up to $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set, such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$:

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}_1} \times \boldsymbol{\mathcal{X}}_{\mathcal{L}_2} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}}} : \quad \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{f}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_1}, \boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \iff \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e})$$

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

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}_2} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}}} : \quad \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \iff \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{g}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e})$$

Take as Ansatz for the mapping $\boldsymbol{g}_{\mathcal{L}_1 \cup \mathcal{L}_2} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_1} \times \boldsymbol{\mathcal{X}}_{\mathcal{L}_2}$:

$$\begin{cases} \boldsymbol{g}_{\mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{L}_1}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) := \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{g}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{g}_{\mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) := \boldsymbol{g}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases}$$

Then for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$:

$$\begin{aligned} \forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_{\mathcal{L}_1} \times \boldsymbol{\mathcal{X}}_{\mathcal{L}_2} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}}} : \quad & \begin{cases} \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{f}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_1}, \boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{L}_1}, \boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \\ \iff & \begin{cases} \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{L}_1}, \boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \\ \iff & \begin{cases} \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{f}_{\mathcal{L}_2}(\boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \\ \iff & \begin{cases} \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{x}_{\mathcal{L}_2}, \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{g}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \\ \iff & \begin{cases} \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1}(\boldsymbol{g}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}), \boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{g}_{\mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \\ \iff & \begin{cases} \boldsymbol{x}_{\mathcal{L}_1} = \boldsymbol{g}_{\mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{L}_1}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{L}_2} = \boldsymbol{g}_{\mathcal{L}_1 \cup \mathcal{L}_2, \mathcal{L}_2}(\boldsymbol{x}_{\mathcal{O}}, \boldsymbol{e}) \end{cases} \end{aligned}$$

which is equivalent to $\mathcal{M}$ being uniquely solvable w.r.t. $\mathcal{L}_1 \cup \mathcal{L}_2$. Similar reasoning shows that the converse holds. Moreover, since the mapping $\boldsymbol{g}_{\mathcal{L}_1}$ is uniquely defined up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set, we have that $\boldsymbol{g}_{\mathcal{L}_2}$ is uniquely defined up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set if and only if $\boldsymbol{g}_{\mathcal{L}_1 \cup \mathcal{L}_2}$ is uniquely defined up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set and hence by the definition of marginalization it follows that $\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)$. □

Note that in the previous Lemma $\mathcal{L}_1$ and $\mathcal{L}_2$ has to be disjoint, since marginalizing first over $\mathcal{L}_1$ gives a maginal SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ with endogenous variables $\mathcal{I} \setminus \mathcal{L}_1$.

This leads to the following commutativity result:

Proposition 39 *Given an SCM $\mathcal{M}$ and disjoint subsets $\mathcal{L}_1, \mathcal{L}_2 \subset \mathcal{I}$. If $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_1 \cup \mathcal{L}_2$, then marginalizing subsequently over $\mathcal{L}_1$ and $\mathcal{L}_2$ commutes, i.e.* $\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1) = \mathrm{marg}(\mathcal{L}_1) \circ \mathrm{marg}(\mathcal{L}_2) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)$.

Proof. This follows from Lemma 38. $\square$

Marginalization is not always defined for all subsets as can be seen in Example 21. There $\mathcal{M}$ is not uniquely solvable with respect to $\{1\}$, nor with respect to $\{2\}$, but it is uniquely solvable with respect to $\{1, 2\}$. Therefore, we cannot marginalize either x_1 or x_2 separately, but we can marginalize over them jointly.

If one starts with an SCM without self-loops, then marginalizing over a subset of endogenous variables may give an SCM that contains self-loops, as we saw in Example 28. However, one can under certain conditions remove the self-loops of an SCM:

Proposition 40 *Given an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with self-loops at endogenous variables $\mathcal{S} \subseteq \mathcal{I}$. Assume for each $i \in \mathcal{S}$ the SCM $\mathcal{M}$ is uniquely solvable w.r.t. i. Let $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ where the causal mechanism $\tilde{\boldsymbol{f}}$ is defined by:*

$$\tilde{f}_i(\boldsymbol{x}, \boldsymbol{e}) := \begin{cases} g_i(\boldsymbol{x}_{\mathcal{I}\setminus\{i\}}, \boldsymbol{e}) & i \in \mathcal{S} \\ f_i(\boldsymbol{x}_{\mathcal{I}\setminus\{i\}}, \boldsymbol{e}) & \text{otherwise}, \end{cases}$$

where g_i is an induced mapping from Lemma 32 and f_i is a representation of the causal mechanism of $\mathcal{M}$ that does not depend on x_i. Then $\mathcal{M} \equiv \tilde{\mathcal{M}}$.

Proof. This follows immediately from Lemma 32 and Definition 5. $\square$

This means that under these conditions one can always find a representation of the SCM without self-loops, and hence these two equivalent representations are in particular observational and interventional equivalent. Moreover, the conditions for removing the self-loops from an SCM are not the same as the conditions that are needed to be able to marginalize over each variable that has a self-loop subsequently, since marginalizing over a variable that has a self-loop may introduce new self-loops, as we saw in Example 28.

For a uniquely solvable SCM, the marginalization does not change the solution space of the observed variables:

Lemma 41 *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L} \subset \mathcal{I}$ of endogenous variables such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$. Then:*

1. *$\mathcal{M}$ is uniquely solvable if and only if $\mathcal{M}_{\text{marg}(\mathcal{L})}$ is uniquely solvable;*
2. *if $\mathcal{M}$ is uniquely solvable, then $\boldsymbol{pr}_{\mathcal{O}} \circ \boldsymbol{g}^{\mathcal{M}} = \boldsymbol{g}^{\mathcal{M}_{\text{marg}(\mathcal{L})}}$ a.e.,* [14] *where $\boldsymbol{pr}_{\mathcal{O}}$ is the projection $\boldsymbol{pr}_{\mathcal{O}} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$.*
3. *if $\mathcal{M}$ is uniquely solvable, then each solution $(\boldsymbol{E}, \boldsymbol{X})$ of $\mathcal{M}$ induces a solution $(\boldsymbol{E}, \boldsymbol{X}_{\mathcal{O}})$ of $\mathcal{M}_{\text{marg}(\mathcal{L})}$;*
4. *if $\mathcal{M}$ is uniquely solvable, then $\mathcal{M}$ and its marginalization $\mathcal{M}_{\text{marg}(\mathcal{L})}$ are observationally equivalent with respect to $\mathcal{I} \setminus \mathcal{L}$.*

Proof.

[14] Here $\boldsymbol{g}^{\mathcal{M}}$ and $\boldsymbol{g}^{\mathcal{M}_{\text{marg}(\mathcal{L})}}$ denote the induced measurable mapping from Lemma 12, where the superscript indicates the respective SCM.

1. As $\mathcal{M}$ is uniquely solvable with respect to $\mathcal{L}$, we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$:

$$\forall_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}_{\mathcal{L}}\times\boldsymbol{\mathcal{X}}_{\mathcal{O}}}: \quad \begin{cases} \boldsymbol{x}_{\mathcal{L}} &= \boldsymbol{f}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{L}},\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{O}} &= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}_{\mathcal{L}},\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}) \end{cases} \iff \begin{cases} \boldsymbol{x}_{\mathcal{L}} &= \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{O}} &= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}),\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e})\,. \end{cases}$$

   Moreover, we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$:

$$\forall_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}_{\mathcal{O}}}: \quad \begin{cases} \boldsymbol{x}_{\mathcal{L}} &= \boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}) \\ \boldsymbol{x}_{\mathcal{O}} &= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}),\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e})\,. \end{cases} \iff \begin{cases} \boldsymbol{x}_{\mathcal{O}} &= \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{g}_{\mathcal{L}}(\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e}),\boldsymbol{x}_{\mathcal{O}},\boldsymbol{e})\,. \end{cases}$$

   Therefore, the marginal SCM is uniquely solvable if and only if the original SCM is uniquely solvable.

2. If we assume in addition $\mathcal{M}$ that is uniquely solvable, then we have an induced mapping $\boldsymbol{g}^{\mathcal{M}}:\boldsymbol{\mathcal{E}}\to\boldsymbol{\mathcal{X}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, $\boldsymbol{g}^{\mathcal{M}}(\boldsymbol{e})=\boldsymbol{f}(\boldsymbol{g}^{\mathcal{M}}(\boldsymbol{e}),\boldsymbol{e})$, and in particular for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, $\boldsymbol{x}_{\mathcal{O}}=\boldsymbol{g}_{\mathcal{O}}(\boldsymbol{e})$ (where $\boldsymbol{g}_{\mathcal{O}}:=\boldsymbol{pr}_{\mathcal{O}}\circ\boldsymbol{g}^{\mathcal{M}}$) is a solution to the above set of equations, and hence $\boldsymbol{pr}_{\mathcal{O}}\circ\boldsymbol{g}^{\mathcal{M}}=\boldsymbol{g}^{\mathcal{M}_{\mathrm{marg}(\mathcal{L})}}$ a.e..

3. This follows directly from 2.

4. By 2. it follows that whenever $\mathcal{M}$ is not uniquely solvable then $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ is not uniquely solvable. Moreover, if $\mathcal{M}$ is uniquely solvable, then the marginal observational distribution over $\boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus\mathcal{L}}$ is given by the push-forward of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ along the composite mapping $\boldsymbol{pr}_{\mathcal{I}\setminus\mathcal{L}}\circ\boldsymbol{g}^{\mathcal{M}}$. But this composite mapping is exactly the mapping $\boldsymbol{g}^{\mathcal{M}_{\mathrm{marg}(\mathcal{L})}}$ from 2. and hence $\mathcal{M}$ and $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ are observationally equivalent with respect to $\mathcal{I}\setminus\mathcal{L}$.

□

An important property of marginalization is that it preserves causal semantics. More precisely, the interventional distributions induced by an SCM are identical to those induced by its marginalization, as long as the interventions do not target the variables that were marginalized over. A simple proof of this result proceeds by showing that the operations of intervening and marginalizing commute.

Lemma 42 *Given an SCM $\mathcal{M}$, a subset $\mathcal{L}\subset\mathcal{I}$ of endogenous variables such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, a subset $I\subseteq\mathcal{I}\setminus\mathcal{L}$ and a value $\boldsymbol{\xi}_I\in\prod_{i\in I}\mathcal{X}_i$. Marginalizing over $\mathcal{L}$ commutes with the perfect intervention $\mathrm{do}(I,\boldsymbol{\xi}_I)$, i.e. $\big(\mathrm{marg}(\mathcal{L})\circ\mathrm{do}(I,\boldsymbol{\xi}_I)\big)(\mathcal{M})=\big(\mathrm{do}(I,\boldsymbol{\xi}_I)\circ\mathrm{marg}(\mathcal{L})\big)(\mathcal{M})$.*

Proof. 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 $\mathcal{L}$ are the same for $\mathcal{M}$ and $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$. □

This leads to our first main result:

Theorem 43 *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L}\subset\mathcal{I}$ of endogenous variables such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$. Then $\mathcal{M}$ and its marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ are interventionally equivalent with respect to $\mathcal{I}\setminus\mathcal{L}$.*

Proof. This follows from Lemma 41 and 42. $\square$

For the acyclic case $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ is always defined for any $\mathcal{L} \subset \mathcal{I}$, since $\mathcal{M}$ is always uniquely solvable w.r.t. $\mathcal{L}$. Moreover, we know from Lemma 18 that acyclicity is preserved under perfect intervention, which makes acyclic SCMs a very interventionally robust and well-studied class of SCMs.

3.1 Linear SCMs

In this section we apply the marginalization operation to linear SCMs, a special class that has seen much attention in the literature (Bollen, 1989).

Definition 44 *We call an SCM $\mathcal{M}$* real *if $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$. We call $\mathcal{M}$* linear*, if moreover each structural equation is a linear combination of its endogenous and exogenous variables:*

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

where $i \in \mathcal{I}$, and $B \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ and $\Gamma \in \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ are matrices, and where $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$ can be any probability measure[15] *on $\mathbb{R}^{\mathcal{J}}$.*

For a subset $\mathcal{L} \subset \mathcal{I}$ we will also use the more shorthand vector-notation:

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

The coefficient B_{ij} is often referred to as the "causal effect" from j to i.

Lemma 45 *Given a linear SCM $\mathcal{M}$ and a subset $\mathcal{L} \subseteq \mathcal{I}$ of endogenous variables. Then $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if the matrix $I - B_{\mathcal{L}\mathcal{L}}$ is invertible. Moreover, if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then the induced mapping $\boldsymbol{g}_{\mathcal{L}} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{L}}$ in Lemma 32 is given for $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$-almost every $\boldsymbol{e}$ by*

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

Proof. $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ implies that for $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$:

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

has a unique solution $\boldsymbol{x}_{\mathcal{L}} \in \mathcal{X}_{\mathcal{L}}$. Hence $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if $I - B_{\mathcal{L}\mathcal{L}}$ is invertible. $\square$

This leads to the following result:

Proposition 46 *Given a linear SCM $\mathcal{M}$ and a subset $\mathcal{L} \subset \mathcal{I}$ of endogenous variables such that $I - B_{\mathcal{L}\mathcal{L}}$ is invertible. Then there exist 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*

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

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

Proof. By the definition of marginalization and Lemma 45 we get for $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$:

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

□

Hence the class of linear SCMs is closed under marginalization. From Theorem 43 we know that $\mathcal{M}$ and its marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ are interventionally equivalent. These results can also be found in Hyttinen et al. (2012).

Remark: In the graphical marginalization operation we had to replace every directed edge $k \to j$ for $j \in \mathcal{O}, k \in \mathcal{L}$ by the set of directed edges $i \to j$ with $i \in \mathcal{O} \cup \mathcal{J}$ whenever there is a directed path $i \to \ell \to \cdots \to k \to j$ where the subpath $\ell \to k$ is a sequence of directed edges between the nodes in $\mathcal{L}$. This substitution of the set of sequences of directed edges in $\mathcal{L}$ is precisely described by the weighted adjacency matrix $(I - B_{\mathcal{L}\mathcal{L}})^{-1}$. In particular, if the spectral radius of $B_{\mathcal{L}\mathcal{L}}$ is less than one, then $(I - B_{\mathcal{L}\mathcal{L}})^{-1} = \sum_{n=0}^{\infty}(B_{\mathcal{L}\mathcal{L}})^n$, i.e. the substitution of the set of sequences is described by the matrix that sums all the weighted adjacency matrices representing paths through latent variables of length n.

4 Exogenous reparametrizations

In the previous section we introduced marginalizations, which map an SCM to another SCM that has only a subset of the endogenous variables of the original SCM, while preserving as much of the causal semantics of the model as possible and leaving the exogenous variables and their distribution invariant. In this section we will introduce another class of mappings, "exogenous reparameterizations", that map an SCM to another SCM with possibly different *exogenous* variables (and in general, a different distribution on those), while preserving all causal semantics of the model and leaving the *endogenous* variables invariant. One can think of a special class of exogenous reparameterizations as "symmetries" that act on the space of (equivalence classes of) SCMs. Therefore, they constitute a natural concept that is interesting in its own right. Furthermore, they will turn out to be useful in the next section, where we study "reductions" of the space of exogenous variables that enable us to derive more parsimonious, but interventionally equivalent, SCMs.

Given an SCM $\mathcal{M}$ with exogenous variable space $\boldsymbol{\mathcal{E}}$ and a mapping $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$, we define an "exogenous reparametrization" as an SCM $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$ of $\mathcal{M}$ that has the same endogenous variables but has exogenous variable space $\tilde{\boldsymbol{\mathcal{E}}}$:

Definition 47 *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 measurable mapping ("reparameterization") $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$, where $\tilde{\boldsymbol{\mathcal{E}}} = \prod_{i \in \tilde{\mathcal{J}}} \tilde{\mathcal{E}}_i$ and* *each domain $\tilde{\mathcal{E}}_i$ is a standard measurable space. An SCM $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})} = \langle \mathcal{I}, \tilde{\mathcal{J}}, \boldsymbol{\mathcal{X}}, \tilde{\boldsymbol{\mathcal{E}}}, \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ is called an* exogenous reparameterization of $\mathcal{M}$ with respect to $\boldsymbol{\phi}$ *if*

- *$\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$ is the push-forward measure of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ under $\boldsymbol{\phi}$, and*
- *for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$:*

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}}: \qquad \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e})).$$

To show that an exogenous reparametrization is defined up to equivalence, we need the following lemma:

Lemma 48 *Let $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$ be a measurable map between to 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 depending on $\tilde{\boldsymbol{e}}$. Then the property $P = \tilde{P} \circ \boldsymbol{\phi}$ holds $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.s. if and only if the property 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 $\boldsymbol{\mathcal{C}} = \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : P(\boldsymbol{e})\}$ contains a Borel 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 Cohn (2013, Proposition 8.2.6), $\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*)$ is analytic. By Cohn (2013, Theorem 8.4.1), there exist Borel 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 $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\mathcal{C}}$, $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*) \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}) \subseteq \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}})\}$. Hence the set $\{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}})\}$ contains a Borel set of $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-measure 1.

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

Proposition 49 *For an SCM $\mathcal{M}$ and a reparameterization $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$: if an exogenous reparameterization of $\mathcal{M}$ with respect to $\boldsymbol{\phi}$ exists,*

1. *it is unique up to equivalence, and*
2. *if $\mathcal{M} \equiv \bar{\mathcal{M}}$ then exogenous reparameterizations $\bar{\mathcal{M}}_{\mathrm{rep}(\boldsymbol{\phi})}$ of $\bar{\mathcal{M}}$ with respect to $\boldsymbol{\phi}$ exist and are equivalent to $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$.*

Proof. For the first statement, it suffices to show that the causal mechanism $\boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}$ is unique up to equivalence. Indeed: if for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}}: \qquad \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e}))$$

and for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}}: \qquad \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e})),$$

then, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$

$$\forall_{\boldsymbol{x} \in \boldsymbol{\mathcal{X}}}: \qquad \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e})) = \tilde{\boldsymbol{f}}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e})).$$

Lemma 48 gives that for $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-almost every $\tilde{\boldsymbol{e}}$

$$\forall_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}}: \qquad \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x},\tilde{\boldsymbol{e}}) = \tilde{\boldsymbol{f}}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x},\tilde{\boldsymbol{e}}).$$

The second statement follows similarly. □

The main result of this section will be that an exogenous reparameterization of an SCM preserves its causal semantics, i.e., it is interventionally equivalent to the original SCM. In general, an exogenous reparameterization may recombine exogenous variables in arbitrary ways, and thereby alter dependencies between exogenous variables. Before we prove several elementary properties of exogenous reparameterizations, we give a few examples.

Example 50 *Consider the uniquely solvable SCM $\mathcal{M}$ from Example 24. The exogenous reparametrization $\boldsymbol{\phi}(\boldsymbol{e}) = \Sigma^{-1}(\boldsymbol{e}-\boldsymbol{\mu})$ with $\Sigma_{12} = \Sigma_{21} = 0$, $\Sigma_{11} = \sigma_1$ and $\Sigma_{22} = \sigma_2$ defines the exogenously reparametrized causal mechanisms*

$$\begin{aligned} f_1(\boldsymbol{x},\boldsymbol{e}) &= \sigma_1 e_1 + \mu_1 \\ f_2(\boldsymbol{x},\boldsymbol{e}) &= \alpha x_1 + \sigma_2 e_2 + \mu_2\,, \end{aligned}$$

with $\mathbb{P}_{\mathbb{R}^2}$ the standard multivariate normal measure on $\mathbb{R}^2$.

Example 51 *Suppose we have an SCM $\mathcal{M} = \langle \mathcal{I},\mathcal{J},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ where $\mathrm{pa}(\mathcal{I})\cap\mathcal{J}$ is not equal to $\mathcal{J}$. This means that there is at least one exogenous variable that is not a parent of any endogenous variable. Then we can consider the exogenous reparametrization with respect to the projection mapping $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})\cap\mathcal{J}}$, which gives the exogenously reparamatrized SCM*

$$\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})} = \langle \mathcal{I}, \mathrm{pa}(\mathcal{I})\cap\mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})\cap\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})\cap\mathcal{J}}}\rangle.$$

This simply disregards a latent exogenous variable that is not a parent of any endogenous variable.

Example 52 *Trivially one can take for an SCM $\mathcal{M} = \langle \mathcal{I},\mathcal{J},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ the exogenous reparametrization with respect to $\boldsymbol{\phi} := \boldsymbol{Id}_{\boldsymbol{\mathcal{E}}} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}$ that maps $\mathcal{M}$ to the SCM $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})} = \langle \mathcal{I},\{1\},\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{E}},\boldsymbol{f},\mathbb{P}_{\boldsymbol{\mathcal{E}}}\rangle$ that has only a single exogenous variable, which combines the components in $\mathcal{E}_j$ to a single tuple in $\boldsymbol{\mathcal{E}}$.*

More motivating examples for exogenous reparameterizations will be given in the next section on reductions.

Exogenous reparameterizations do not always exist: for given SCM $\mathcal{M}$ and a measurable reparametrization $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$ it may not be possible to construct a measurable mapping $\boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})} : \boldsymbol{\mathcal{X}} \times \tilde{\boldsymbol{\mathcal{E}}} \to \boldsymbol{\mathcal{X}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x}\in\boldsymbol{\mathcal{X}}$, $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{e}) = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x},\boldsymbol{\phi}(\boldsymbol{e}))$ holds. As an example, consider taking $\tilde{\boldsymbol{\mathcal{E}}} = \{0\}$, i.e., the whole exogenous space is mapped to a single point. Only if the original SCM is deterministic, an exogenous reparameterization with respect to (the only possible) $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$ exists.

A sufficient condition for the existence of an exogenous reparameterization is that $\boldsymbol{\phi}$ has a measurable right-sided inverse $\boldsymbol{\pi}$ (i.e., $\boldsymbol{\phi}\circ\boldsymbol{\pi} = \mathbf{Id}_{\tilde{\boldsymbol{\mathcal{E}}}}$).

When $\boldsymbol{\phi}$ has such a right-sided inverse $\boldsymbol{\pi}$, the exogenously reparametrized causal mechanism can be defined as $\boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})} := \boldsymbol{f} \circ (\mathrm{Id}_{\boldsymbol{\mathcal{X}}}, \boldsymbol{\pi})$.

In particular, this construction is applicable when $\boldsymbol{\phi}$ is an isomorphism (i.e., $\boldsymbol{\phi}$ is a bijection, $\boldsymbol{\phi}$ is measurable, and $\boldsymbol{\phi}^{-1}$ is measurable).

Proposition 53 *1. Exogenous reparametrization preserves composition, i.e.,* $\mathrm{rep}(\boldsymbol{\phi}_2) \circ \mathrm{rep}(\boldsymbol{\phi}_1) = \mathrm{rep}(\boldsymbol{\phi}_2 \circ \boldsymbol{\phi}_1)$ *whenever defined.*

2. If $\boldsymbol{\phi}$ is an isomorphism, $\mathrm{rep}(\boldsymbol{\phi})$ *is defined, and* $\mathrm{rep}(\boldsymbol{\phi}^{-1}) = \mathrm{rep}(\boldsymbol{\phi})^{-1}$.

Proof. For the first statement: for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$:

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) &= \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi}_1)}(\boldsymbol{x}, \boldsymbol{\phi}_1(\boldsymbol{e})) \\ &= (\boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi}_1)})_{\mathrm{rep}(\boldsymbol{\phi}_2)}(\boldsymbol{x}, \boldsymbol{\phi}_2 \circ \boldsymbol{\phi}_1(\boldsymbol{e}))\end{aligned}$$

which defines $\boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi}_2 \circ \boldsymbol{\phi}_1)}$ up to equivalence. The second statement follows from the first statement. □

Therefore, exogenous reparameterizations with respect to isomorphisms form a group action on the space of (equivalence classes of) SCMs. In the remainder of this section, we will see that they keep the causal semantics invariant. Therefore, we can think of this class of exogenous reparameterizations as symmetries of the space of (equivalence classes of) SCMs.

Similarly to marginalizations, exogenous reparametrizations have the property that they preserve the causal semantics. The proof proceeds in a similar series of steps as for marginalizations: first we prove that an exogenous reparameterization is observationally equivalent, then we show that exogenous reparameterizations commute with interventions, and finally we conclude that exogenous reparameterizations are interventionally equivalent.

Lemma 54 *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 measurable mapping* $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$ *resulting in the exogenously reparametrized SCM* $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})} = \langle \mathcal{I}, \tilde{\mathcal{J}}, \boldsymbol{\mathcal{X}}, \tilde{\boldsymbol{\mathcal{E}}}, \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$*. Then:*

1. *$\mathcal{M}$ is uniquely solvable if and only if $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$ is uniquely solvable;*
2. *if $\mathcal{M}$ is uniquely solvable, then $\mathcal{M}$ and $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$ are observationally equivalent;*
3. *if $\mathcal{M}$ is uniquely solvable, each solution $(\boldsymbol{E}, \boldsymbol{X})$ of $\mathcal{M}$ induces a solution $(\boldsymbol{\phi}(\boldsymbol{E}), \boldsymbol{X})$ of $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$.*

Proof.

1. Note that the set

$$\boldsymbol{\mathcal{E}}_0 := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \ : \ \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e}) = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e}))\}$$

contains $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$. We start by assuming that $\mathcal{M}$ is uniquely solvable, the same holds for the set

$$\boldsymbol{\mathcal{E}}_1 := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \exists! \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \ : \ \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})\}.$$

Therefore, their intersection $\boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{E}}_1$ contains a Borel set $\boldsymbol{\mathcal{E}}^*$ with $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measure 1. Moreover $\boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{E}}_1$ is a subset of

$$\boldsymbol{\mathcal{E}}_\cap := \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : \exists! \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \ : \ \boldsymbol{x} = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \boldsymbol{\phi}(\boldsymbol{e}))\}$$

and thus $\boldsymbol{\mathcal{E}}^* \subseteq \boldsymbol{\mathcal{E}}_\cap$. By Lemma 48 if follows that the set

$$\tilde{\boldsymbol{\mathcal{E}}}_1 := \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \exists! \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \ : \ \boldsymbol{x} = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}, \tilde{\boldsymbol{e}})\}.$$

contains a Borel set with $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-measure 1 and hence $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$ is uniquely solvable.

For the converse, suppose $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$ is uniquely solvable. Then $\tilde{\boldsymbol{\mathcal{E}}}_1$ contains a Borel set of $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-measure 1. By Lemma 48 it follows that $\boldsymbol{\mathcal{E}}_\cap$ contains a Borel set of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measure 1. Since $\boldsymbol{\mathcal{E}}_0$ contains almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ we have that their intersection $\boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{E}}_\cap$ contains a Borel set of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measure 1. Now since $\boldsymbol{\mathcal{E}}_0 \cap \boldsymbol{\mathcal{E}}_\cap \subseteq \boldsymbol{\mathcal{E}}_1$, we have that $\mathcal{M}$ is uniquely solvable.

2. We now show that $\mathcal{M}$ and $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$ are observationally equivalent, assuming that $\mathcal{M}$ is uniquely solvable. Let $\boldsymbol{g} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ be the induced mapping (as in Lemma 12), and let $\boldsymbol{g}_{\mathrm{rep}(\boldsymbol{\phi})} : \tilde{\boldsymbol{\mathcal{E}}} \to \boldsymbol{\mathcal{X}}$ be the corresponding mapping for $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$. Note that $\boldsymbol{g}_{\mathrm{rep}(\boldsymbol{\phi})} \circ \boldsymbol{\phi} = \boldsymbol{g}$ a.e., as for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$

$$\boldsymbol{g}(\boldsymbol{e}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{e}), \boldsymbol{e}) = \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{g}(\boldsymbol{e}), \boldsymbol{\phi}(\boldsymbol{e})) = \boldsymbol{g}_{\mathrm{rep}(\phi)}(\boldsymbol{\phi}(\boldsymbol{e})).$$

The observational distribution of $\mathcal{M}$ is the push-forward of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ under $\boldsymbol{g}$. As $\boldsymbol{g} = \boldsymbol{g}_{\mathrm{rep}(\boldsymbol{\phi})} \circ \boldsymbol{\phi}$ a.e., this equals the push-forward under $\boldsymbol{g}_{\mathrm{rep}(\boldsymbol{\phi})}$ of the push-forward under $\boldsymbol{\phi}$ of $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$. As the latter equals $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$, this equals the observational distribution of $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$.

3. Let $(\boldsymbol{E}, \boldsymbol{X})$ be a solution of a uniquely solvable $\mathcal{M}$ and assume some background probability space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, \mathbb{P})$. Then the following diagram commutes

$$\begin{tikzcd} \Omega \arrow[r, "\boldsymbol{E}"] \arrow[dr, "\boldsymbol{X}"'] & \boldsymbol{\mathcal{E}} \arrow[r, "\boldsymbol{\phi}"] \arrow[d, "\boldsymbol{g}"] & \tilde{\boldsymbol{\mathcal{E}}} \arrow[dl, "\tilde{\boldsymbol{g}}"] \\ & \boldsymbol{\mathcal{X}} & \end{tikzcd}$$

almost surely, and hence the outer triangle commutes also almost surely. This means that $(\boldsymbol{\phi}(\boldsymbol{E}), \boldsymbol{X})$ is a solution of $\mathcal{M}_{\mathrm{rep}(\boldsymbol{\phi})}$.

□

The following observation is trivial, yet crucial:

Lemma 55 *Exogenous reparametrization and perfect intervention of an SCM commute.*

Proof. This follows directly from Definition 15 and 47. □

It directly leads to our second main result:

Theorem 56 *An exogenously reparametrized SCM is interventionally equivalent to the original SCM.*

Proof. This follows directly from Lemma 54 and 55. □

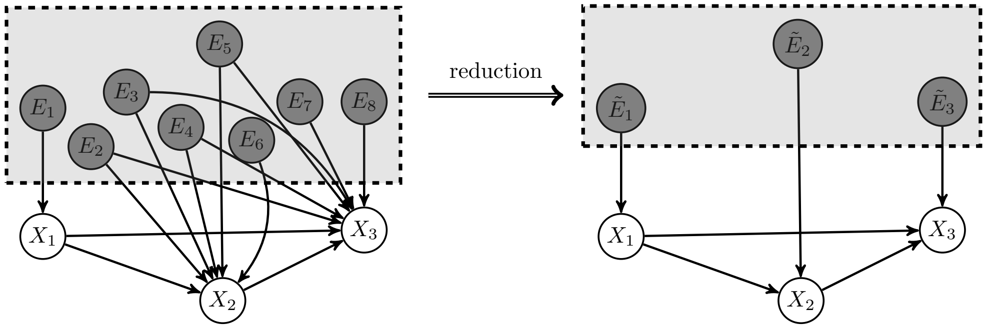

Figure 6: Example of the augmented graphs of an SCM (left) and its reduction (right).

5 Reductions

Whereas marginalizations allow us to focus our causal modeling efforts on a part of a system while ignoring the rest of the system, they operate only on the endogenous part of the system, leaving the exogenous part invariant. This may lead to models that have only a few endogenous variables but a very large number of exogenous variables, which is not very practical. The central question of this section is whether one can represent the influence of the surroundings of the system of interest in a more parsimonious way (see also Figure 6).

We will introduce a general concept of such a "reduction" of the exogenous space. Our main motivation for this notion of reduction is that it extends special cases of interest that are already known in the literature (in particular, reductions of linear SCMs and of real Markovian SCMs). It turns out that such general reductions always exist and can be obtained by an exogenous reparameterization. However, reduced SCMs obtained in that way will be generally impossible to learn from data, as their causal mechanisms are typically very wild. We argue that additional "smoothness" assumptions are required when the goal is to arrive at a reduced SCM that can be estimated from data. We study the existence of a certain class of "nice" reductions that are smoother, but still capture the special cases of reductions of linear SCMs and of real Markovian SCMs. The main, and somewhat disappointing, result of this section will be a counterexample that shows that such "nice" reductions do not exist in general for nonlinear SCMs with confounding latent variables.

5.1 Reductions of linear SCMs

Before we give a general definition of a reduction, we will look at one of the primary motivating examples, the linear case. For linear SCMs, we saw in Proposition 46 that marginalizing over a subset of endogenous variables leads to a linear model with fewer endogenous variables, but with the same number of exogenous variables. If the dimension $|\mathcal{J}|$ of the space of exogenous variables is larger than the dimension $|\mathcal{O}|$ of the space of the resulting endogenous variables, we would like to obtain a more parsimonious representation with fewer exogenous variables. In the linear SCM setting, this can be achieved by a certain exogenous reparametrization, as the following example shows:

Example 57 *In the same context as that of Proposition 46, consider the exogenous reparametrization of the marginalized SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ with respect to the reparametrization $\boldsymbol{\phi} : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{O}}$ defined by:*

$$\boldsymbol{\phi}(\boldsymbol{e}) = [B_{\mathcal{O}\mathcal{L}}(I - B_{\mathcal{L}\mathcal{L}})^{-1}\Gamma_{\mathcal{L}\mathcal{J}} + \Gamma_{\mathcal{O}\mathcal{J}}]\boldsymbol{e}. \tag{4}$$

This yields the reduction

$$\mathcal{M}_{\mathrm{marg}(\mathcal{L}),\mathrm{rep}(\boldsymbol{\phi})} = \langle \mathcal{O}, \mathcal{O}, \mathbb{R}^{\mathcal{O}}, \mathbb{R}^{\mathcal{O}}, \tilde{\boldsymbol{f}}_{\mathrm{rep}(\boldsymbol{\phi})}, \mathbb{P}_{\mathbb{R}^{\mathcal{O}}} \rangle$$

of $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$, where the exogenously reparametrized marginal causal mechanism $\tilde{\boldsymbol{f}}_{\mathrm{rep}(\boldsymbol{\phi})} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{O}} \to \mathbb{R}^{\mathcal{O}}$ is given by

$$\tilde{\boldsymbol{f}}_{\mathrm{rep}(\boldsymbol{\phi})}(\boldsymbol{x}_{\mathcal{O}}, \tilde{\boldsymbol{e}}) = [B_{\mathcal{O}\mathcal{O}} + B_{\mathcal{O}\mathcal{L}}(I - B_{\mathcal{L}\mathcal{L}})^{-1}B_{\mathcal{L}\mathcal{O}}]\boldsymbol{x}_{\mathcal{O}} + \tilde{\boldsymbol{e}}$$

and $\mathbb{P}_{\mathbb{R}^{\mathcal{O}}}$ is given by the push-forward of $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$ along the reparametrization $\boldsymbol{\phi}$. It follows from Theorem 56 that the reduced SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L}),\mathrm{rep}(\boldsymbol{\phi})}$ is interventionally equivalent to the marginalized SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ (which is interventionally equivalent to the original SCM $\mathcal{M}$ by Theorem 43).

If all exogenous variables $\boldsymbol{E}$ of a solution of the original linear SCM $\mathcal{M}$ are jointly independent, then the diagonal covariance matrix $\Sigma(\boldsymbol{E})$ of $\boldsymbol{E}$ is transformed under (4) to

$$\Sigma(\tilde{\boldsymbol{E}}) = \boldsymbol{\Phi}\Sigma(\boldsymbol{E})\boldsymbol{\Phi}^T,$$

where $\boldsymbol{\Phi} = B_{\mathcal{O}\mathcal{L}}(I - B_{\mathcal{L}\mathcal{L}})^{-1}\Gamma_{\mathcal{L}\mathcal{J}} + \Gamma_{\mathcal{O}\mathcal{J}}$. In general, $\Sigma(\tilde{\boldsymbol{E}})$ may not be diagonal, even if $\Sigma(\boldsymbol{E})$ is diagonal.

This example shows that in the linear case, a reduction as illustrated in Figure 6 can be achieved via a certain exogenous reparametrization (4). Whereas the dimensionality of the original exogenous space could be very large (i.e., $|\mathcal{L}| \gg |\mathcal{O}|$), the reduced SCM has only a single real-valued exogenous variable for each endogenous variable, leading to a more parsimonious representation. The example also illustrates that the reduced exogenous variables may no longer be jointly independent, even when the original exogenous variables were. The covariances between exogenous variables that are introduced reflect underlying confounding between variables (intuitively, the existence of latent common causes). Generally, an exogenous reparameterization that combines several exogenous variables into new ones will introduce additional dependencies between exogenous variables.

Our results for linear SCMs are strongly related and in fact inspired by the work of Hyttinen et al. (2012). In that work, only linear SCMs are considered, and the authors consider the exogenous reparametrization that we described above to be an integral part of the marginalization operation. We decided to decouple the exogenous reparameterization from the marginalization operation, as this makes it easier to deal with the general (nonlinear) case. In addition, our results justify the implicit assumption made by Hyttinen et al. (2012) that one may assume the number of exogenous variables to be equal to the number of endogenous variables (for linear SCMs) without loss of generality.

5.2 General notion of reductions

After this motivating example, we now introduce the central concept of this section:

Definition 58 *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$ be a real SCM. We call an SCM $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{K}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{K}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^{\mathcal{K}}} \rangle$* a reduction *of $\mathcal{M}$ if $\tilde{\mathcal{M}}$ is interventionally equivalent to $\mathcal{M}$, and $|\mathcal{K}| < |\mathcal{J}|$.*

In Example 57, we saw that linear SCMs can be reduced such that only a single one-dimensional exogenous variable is necessary for each endogenous variable. Our following example is a slight generalization of this idea to nonlinear SCMs:

Example 59 *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$, where for each $i \in \mathcal{I}$:*

$$f_i(\boldsymbol{x}, \boldsymbol{e}) = \psi_i(\boldsymbol{x}, h_i(\boldsymbol{e})),$$

with $h_i : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}$ and $\psi_i : \mathbb{R}^{\mathcal{I}} \times \mathbb{R} \to \mathbb{R}$ measurable functions. Similarly to linear SCMs, if $|\mathcal{J}| > |\mathcal{I}|$, these SCMs admit a reduction that can be obtained as an exogenous reparametrization. Indeed, the exogenous reparametrization $\boldsymbol{\phi} : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{I}}$ defined by $\phi_i(\boldsymbol{e}) := h_i(\boldsymbol{e})$ yields the exogenously reparametrized SCM $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{I}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{I}}, \boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})}, \mathbb{P}_{\mathbb{R}^{\mathcal{I}}} \rangle$ with causal mechanism $\boldsymbol{f}_{\mathrm{rep}(\boldsymbol{\phi})} : \mathbb{R}^{\mathcal{I}} \times \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{I}}$ given by

$$f_{\mathrm{rep}(\boldsymbol{\phi}),i}(\boldsymbol{x}, \tilde{\boldsymbol{e}}) := \psi_i(\boldsymbol{x}, \tilde{e}_i).$$

5.3 Reductions of real Markovian SCMs

The next class of examples of reductions (of real Markovian SCMs) has been described before in the literature (see, e.g., Peters et al. (2014); Mooij et al. (2016)) and further motivates the general definition we gave in the previous subsection. An SCM is a real Markovian SCM if it is an acyclic, real SCM, has jointly independent exogenous variables, i.e., $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j \in \mathcal{J}} \mathbb{P}_{\boldsymbol{\mathcal{E}}_j}$, with at most one child per exogenous variable.

Before we can give the example, we need a well-known result on the interventional distributions of Markovian SCMs. It states that for Markovian SCMs, the interventional distribution of a variable when intervening on its direct causes equals the conditional observational distribution of a variable given its direct causes:

Lemma 60 *Let $\mathcal{M}$ be a Markovian SCM. For any $i \in \mathcal{I}$, the family of interventional distributions $\mathbb{P}(X_i \,|\, \mathrm{do}(\mathrm{pa}(i) \cap \mathcal{I}, \boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}))$ is a version of the regular conditional probability $\mathbb{P}(X_i \,|\, \boldsymbol{X}_{\mathrm{pa}(i)\cap\mathcal{I}} = \boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}})$.*

Proof. Let $(\boldsymbol{X}, \boldsymbol{E})$ be a solution of $\mathcal{M}$. Then, for any measurable set A:

$$\begin{aligned}
\mathbb{P}(X_i \in A \,|\, \boldsymbol{X}_{\mathrm{pa}(i)\cap\mathcal{I}} = \boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}) &= \mathbb{P}(f_i(\boldsymbol{X}_{\mathrm{pa}(i)\cap\mathcal{I}}, \boldsymbol{E}_{\mathrm{pa}(i)\cap\mathcal{I}}) \in A \,|\, \boldsymbol{X}_{\mathrm{pa}(i)\cap\mathcal{I}} = \boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}) \\
&= \mathbb{P}(f_i(\boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}, \boldsymbol{E}_{\mathrm{pa}(i)\cap\mathcal{I}}) \in A \,|\, \boldsymbol{X}_{\mathrm{pa}(i)\cap\mathcal{I}} = \boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}) \\
&= \mathbb{P}(f_i(\boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}, \boldsymbol{E}_{\mathrm{pa}(i)\cap\mathcal{I}}) \in A) \\
&= \mathbb{P}(X_i \in A \,|\, \mathrm{do}(\mathrm{pa}(i) \cap \mathcal{I}, \boldsymbol{x}_{\mathrm{pa}(i)\cap\mathcal{I}}))
\end{aligned}$$

a.s., where we used the independence of $\boldsymbol{E}_{\mathrm{pa}(i)\cap\mathcal{I}}$ and $\boldsymbol{X}_{\mathrm{pa}(i)\cap\mathcal{I}}$. □

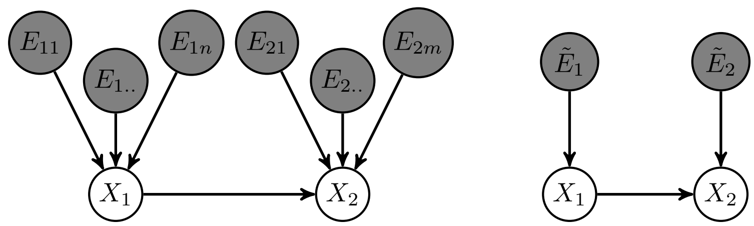

Figure 7: Example illustrating the augmented graph of a possible SCM in Example 61 (left) and the augmented graph of its reduction (right).

We can now state our next example of a class of reductions. Rather than proving this in general that real Markovian SCMs admit reductions, we only provide an example for the simplest case.

Example 61 *Given a Markovian SCM of the form* $\mathcal{M} = \langle \mathbf{2}, \mathcal{J}, \mathbb{R}^2, \mathbb{R}^{n+m}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{n+m}} \rangle$, *where* $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$ *with* $|\mathcal{J}_1| = n$ *and* $|\mathcal{J}_2| = m$, *such that* f_1 *only depends on* $\boldsymbol{e}_{\mathcal{J}_1} \in \mathbb{R}^n$ *and* f_2 *depends on* x_1 *and* $\boldsymbol{e}_{\mathcal{J}_2} \in \mathbb{R}^m$ *(see also Figure 7). If* $n + m > 2$, *we can construct a reduction* $\tilde{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \mathbb{R}^2, \mathbb{R}^2, \tilde{\boldsymbol{f}}, \lambda_{(0,1)^2} \rangle$ *of* $\mathcal{M}$, *where* $\lambda_{(0,1)^2}$ *is a uniform distribution on* $(0,1)^2$, *as follows.*

Let $\big((X_1, X_2), (\boldsymbol{E}_{\mathcal{J}_1}, \boldsymbol{E}_{\mathcal{J}_2})\big)$ *be a solution of* $\mathcal{M}$. *Let*

$$\tilde{f}_1 : (0,1) \to \mathbb{R} : \tilde{e}_1 \mapsto (F_{X_1})^{-1}(\tilde{e}_1).$$

where F_{X_1} *is the distribution function of* X_1. *Then for* $\tilde{E}_1 \sim U(0,1)$ *we have* $\tilde{f}_1(\tilde{E}_1) \sim f_1(\boldsymbol{E}_{\mathcal{J}_1}) \sim X_1$. *Let*

$$\tilde{f}_2 : \mathbb{R} \times (0,1) \to \mathbb{R} : (x_1, \tilde{e}_2) \mapsto (F_{f_2(x_1, \boldsymbol{E}_{\mathcal{J}_2})})^{-1}(\tilde{e}_2).$$

Then for $\tilde{E}_2 \sim U(0,1)$ *we have that for all* x_1, $\tilde{f}_2(x_1, \tilde{E}_2) \sim f_2(x_1, \boldsymbol{E}_{\mathcal{J}_2})$. *This means that for any perfect intervention* $\mathrm{do}(1, x_1)$, *the interventional distributions of the two SCMs are identical. Applying Lemma 60, we obtain that the conditional distribution of* X_2 *given* X_1 *is correctly captured by* $\tilde{\mathcal{M}}$. *This implies that* $\tilde{\mathcal{M}}$ *and* $\mathcal{M}$ *are observationally equivalent. In addition, for any intervention* $\mathrm{do}(2, x_2)$ *the interventional distributions of the two SCMs are identical, as these are all identical to the observational distribution of* X_1. *Therefore,* $\tilde{\mathcal{M}}$ *is even interventionally equivalent to* $\mathcal{M}$.[16]

This example also illustrates that reductions can also be obtained in other ways than using an exogenous reparameterization.

5.4 Existence of reductions

After having seen these motivating examples of reductions, we will now show that reductions exist in general. The heart of the proof is the nontrivial classical result (Kechris, 1995; Cohn, 2013) that classifies all standard measurable spaces up to isomorphisms:

[16] Similar but weaker statement than Example 61 can be found in Peters et al. (2014); Mooij et al. (2016).

Lemma 62 *Two standard measurable spaces are isomorphic if and only if they have the same cardinality. Furthermore, the cardinality of each uncountable standard measurable space is that of the continuum.*

It follows that a standard measurable space is either isomorphic to $\mathbb{R}$, the natural numbers $\mathbb{N}$, the finite sets $\{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$, or the empty set $\emptyset$. In particular, the cardinality of every uncountable standard measurable space is isomorphic to the cardinality of the continuum.[17] For example, there exists an isomorphism $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^m$ for every $n, m \in \mathbb{N}^*$. Hence, using this powerful result and Theorem 56, it follows immediately that:

Corollary 63 *Every real SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$ can be exogenously reparametrized to a real SCM with $\tilde{\boldsymbol{\mathcal{E}}} = \mathbb{R}^d$, i.e., with an arbitrary number d of real-valued exogenous variables. In particular, reductions of real SCMs with $|\mathcal{J}| > |\mathcal{I}|$ exist.*

Note that exogenously reparametrizing the exogenous variables to a single real-valued exogenous variable makes all endogenous variables (that have exogenous direct causes) confounded. However, the set of conditional independencies that holds in the observational distribution will not change under this exogenous reparametrization, since exogenous reparametrization lead to observationally (even interventionally) equivalent SCMs.

In particular, the reductions of linear SCMs described in Example 57, can be reduced even further to a single one-dimensional exogenous variable. This strong reduction comes at a price, though, since the exogenously reparametrized SCM will not be linear anymore. Even worse, the measure-theoretic isomorphisms $\mathbb{R}^d \to \mathbb{R}$ are very "wild" functions, and the reduction that is obtained via an exogenous reparametrization with respect to such an isomorphism will inherit this wildness. In particular, this will make it practically impossible to learn the causal mechanism of such a reduction from data. This suggests that we need to impose some additional conditions on reductions.

5.5 Smoother reductions

Examples 57 and 61 suggest that smoother reductions than the ones from Corollary 63 may exist. For example, the reductions in Example 57 of linear SCMs are linear, and therefore very smooth, SCMs. As a consequence, when learning a linear SCM from data, one can assume without loss of generality that the SCM has a single one-dimensional exogenous variable for each endogenous variable. The model is then completely specified by the parameters of the linear structural equations and by the joint distribution $\mathbb{P}_{\mathbb{R}^{\mathcal{I}}}$ of the (reduced) exogenous variables. This observation raises the question whether smoother reductions than the ones from Corollary 63 exist also in the nonlinear case, which would then provide an elegant approach to learning nonlinear SCMs with latent variables from data: no

[17] The continuum hypothesis, stating that there is no set whose cardinality is strictly between that of the natural numbers and the continuum, is independent of the axioms of set theory, including the axiom of choice. For standard measurable spaces Lemma 62 follows from the axioms of set theory without relying on the continuum hypothesis. For a more comprehensive discussion, see Dudley (2002, Section 13.1).

matter how many exogenous variables the "true" SCM has, one could assume without loss of generality that is has as many exogenous variables as endogenous variables, reducing the estimation of the complex true SCM to the estimation of the more parsimonious reduced SCM. The existence of such smoother reductions is the topic of this subsection.

We will need the following elementary result in probability theory:

Lemma 64 *Let X be a real-valued random variable, F_X its distribution function and let the quantile function F_X^{-1} be its generalised inverse. Then $F_X^{-1}\big(F_X(X)\big) = X$ a.s..*

Proof. Note $F_X^{-1}\big(F_X(x)\big) \neq x$ if and and only if x is in the interior or right end point of an interval on which F_X is constant. For every interval $I = [a, b)$ or $I = [a, b]$ on which F_X is constant corresponds with a discontinuity of F_X^{-1} and we have $\mathbb{P}(X \in I) = 0$ for every such interval. Moreover, since the quantile function F_X^{-1} is monotonically increasing, it can only have countably many discontinuities. As the countable union of sets of measure 0 has measure 0, $\mathbb{P}\big(\{F_X^{-1}(F_X(X)) \neq X\}\big) = 0$. $\square$

Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and let $(\boldsymbol{E}, \boldsymbol{X})$ be a solution. For $i \in \mathcal{I}$ and $\boldsymbol{\xi}_{\setminus i} \in \boldsymbol{\mathcal{X}}_{\setminus i}$, define the function

$$F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{\xi}_{\setminus i})}} : \mathcal{X}_i \to [0, 1] : x_i \mapsto \mathbb{P}(f_i(\boldsymbol{\xi}_{\setminus i}, \boldsymbol{E}) \leq x_i),$$

i.e., the distribution function of endogenous variable i after the perfect intervention $\mathrm{do}(\setminus i, \boldsymbol{\xi}_{\setminus i})$ on all endogenous variables except i itself. These functions give rise to strong constraints on the possible forms of the causal mechanism:

Lemma 65 *Let $\bar{\mathcal{M}} = \langle \mathcal{I}, \mathcal{I}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{I}}, \bar{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^{\mathcal{I}}} \rangle$ be interventionally equivalent to $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$. If*

(i) $\bar{\mathcal{M}}$ has no self-loops, and

(ii) for each $i \in \mathcal{I}$, for all $\boldsymbol{x} \in \mathbb{R}^{\mathcal{I}}$, the function $\bar{e}_i \mapsto \bar{f}_i(\boldsymbol{x}_{\setminus i}, \bar{e}_i)$ is strictly monotonically increasing,

then for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, for $\mathbb{P}_{\mathbb{R}^{\mathcal{I}}}$-almost every $\bar{\boldsymbol{e}}$:

$$\bar{f}_i(\boldsymbol{x}, \bar{\boldsymbol{e}}) = \left(F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{x}_{\setminus i})}}\right)^{-1} \big(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i)\big) \tag{5}$$

where $F_{\bar{E}_i}^{\bar{\mathcal{M}}}$ is the distribution function of $\bar{E}_i$.

Proof. By definition, and by the assumption that $\bar{e}_i \mapsto \bar{f}_i(\boldsymbol{\xi}_{\setminus i}, \bar{e}_i)$ is strictly monotonically increasing for all $\boldsymbol{\xi}_{\setminus i}$, we have for all $\boldsymbol{\xi}_{\setminus i}$ and all $\bar{e}_i$:

$$\begin{aligned} F_{X_i}^{\bar{\mathcal{M}}_{\mathrm{do}(\setminus i, \boldsymbol{\xi}_{\setminus i})}}\big(\bar{f}_i(\boldsymbol{\xi}_{\setminus i}, \bar{e}_i)\big) &= \mathbb{P}(\bar{f}_i(\boldsymbol{\xi}_{\setminus i}, \bar{E}_i) \leq \bar{f}_i(\boldsymbol{\xi}_{\setminus i}, \bar{e}_i)) \\ &= \mathbb{P}(\bar{E}_i \leq \bar{e}_i) \\ &= F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i). \end{aligned} \tag{6}$$

Applying Lemma 64 to the random variable $f_i(\boldsymbol{\xi}_{\setminus i}, \bar{E}_i)$, we obtain for $\mathbb{P}_{\bar{E}_i}$-almost every $\bar{e}_i$

$$\bar{f}_i(\boldsymbol{\xi}_{\setminus i}, \bar{e}_i) = \left(F_{X_i}^{\bar{\mathcal{M}}_{\mathrm{do}(\setminus i, \boldsymbol{\xi}_{\setminus i})}}\right)^{-1} \big(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i)\big). \tag{7}$$

Because of the interventional equivalence of $\mathcal{M}$ and $\bar{\mathcal{M}}$, for all $\boldsymbol{\xi}_{\setminus i}$:

$$F_{X_i}^{\bar{\mathcal{M}}_{\mathrm{do}(\setminus i, \boldsymbol{\xi}_{\setminus i})}} = F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{\xi}_{\setminus i})}}$$

and hence (5) follows. $\square$

However, these conditions are not strong enough to determine the causal mechanism up to equivalence (note the ordering of the quantifiers in (5) is the wrong way around). The following conditions do suffice:

Proposition 66 *If the conditions of Lemma 65 hold, and if in addition, for each $i \in \mathcal{I}$:*

(iii) $\bar{E}_i$ *has a continuous density;*

(iv) $\bar{f}_i$ *is continuous;*

(v) for each $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\left(F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{x}_{\setminus i})}}\right)^{-1}$ *is continuous,*

then $\bar{f}_i$ *is determined up to equivalence, i.e., for* $\mathbb{P}_{\mathbb{R}^{\mathcal{I}}}$*-almost every* $\bar{\boldsymbol{e}}$*:*

$$\forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}} : \qquad \bar{f}_i(\boldsymbol{x}, \bar{\boldsymbol{e}}) = \left(F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{x}_{\setminus i})}}\right)^{-1} \big(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i)\big). \tag{8}$$

Proof. If for two continuous functions $g, \tilde{g} : \mathbb{R} \to \mathbb{R}$ we have that for $\mathbb{P}_{\bar{E}_i}$-almost every $\bar{e}_i$, $g(\bar{e}_i) = \tilde{g}(\bar{e}_i)$, and if we assume that $\bar{E}_i$ has a continuous strictly positive density, then for all $\bar{e}_i \in \mathbb{R}$, $g(\bar{e}_i) = \tilde{g}(\bar{e}_i)$. If the density of $\bar{E}_i$ is continuous but not everywhere positive, then for all $\bar{e}_i$ where the density of $\bar{E}_i$ is strictly positive we have $g(\bar{e}_i) = \tilde{g}(\bar{e}_i)$. Therefore, if $\bar{E}_i$ has a continuous density, and if $\bar{f}_i$ is continuous, then for each $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ for which $\left(F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{x}_{\setminus i})}}\right)^{-1}$ is continuous, (5) holds for all $\bar{\boldsymbol{e}}$ where the density of $\bar{E}_i$ is strictly positive. $\square$

This proposition will not be used later, but it seems interesting in its on right, as it gives conditions under which a reduction is determined up to equivalence. If one can assume these conditions to hold, the reduction is in principle identifiable from the family of all interventional distributions, up to the arbitrary marginal distributions of the exogenous variables.

We are now ready to define a refinement of the notion of reduction. This definition may appear ad-hoc, but it is justified by the next Lemma, which shows that these conditions are sufficient to derive strong identifiability results.

Definition 67 *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$ *be a real SCM. We call an SCM* $\bar{\mathcal{M}} = \langle \mathcal{I}, \mathcal{I}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{I}}, \bar{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^{\mathcal{I}}} \rangle$ *a* nice reduction *of* $\mathcal{M}$ *if*

1. $\bar{\mathcal{M}}$ *is a reduction of* $\mathcal{M}$*, and*
2. $\bar{\mathcal{M}}$ *has no self-loops, and*
3. *for each* $i \in \mathcal{I}$*, for all* $\boldsymbol{x} \in \mathbb{R}^{\mathcal{I}}$*, the function* $\bar{e}_i \mapsto \bar{f}_i(\boldsymbol{x}_{\setminus i}, \bar{e}_i)$ *is strictly monotonically increasing, and*
4. *the exogenous variable* $\bar{E}_i$ *has a continuous distribution function* $F_{\bar{E}_i}^{\bar{\mathcal{M}}}$ *(marginally).*

The following lemma shows that if a nice reduction exists, then there also exists a nice reduction for which each exogenous variable has a standard-normal marginal distribution:

Lemma 68 *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$ be a real SCM. If $\mathcal{M}$ admits a nice reduction, then it admits a nice reduction $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{I}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{I}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^{\mathcal{I}}} \rangle$ for which each exogenous variable $\tilde{E}_i$ has a standard-normal marginal distribution. Its causal mechanism is to a large extent (but not necessarily up to equivalence) determined by $\mathcal{M}$, i.e., for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, for $\mathbb{P}_{\mathbb{R}^{\mathcal{I}}}$-almost every $\tilde{\boldsymbol{e}}$:*

$$\tilde{f}_i(\boldsymbol{x}, \tilde{\boldsymbol{e}}) = \left(F_{X_i}^{\mathcal{M}_{\mathrm{do}(\setminus i, \boldsymbol{x}_{\setminus i})}}\right)^{-1}\big(\Phi(\tilde{e}_i)\big) \tag{9}$$

where Φ is the distribution function of a standard-normal distribution over $\mathbb{R}$.

Proof. Let $\bar{\mathcal{M}} = \langle \mathcal{I}, \mathcal{I}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{I}}, \bar{\boldsymbol{f}}, \mathbb{P}_{\mathbb{R}^{\mathcal{I}}} \rangle$ be a nice reduction of $\mathcal{M}$. Consider the exogenous reparameterization $\tilde{\mathcal{M}}$ of $\bar{\mathcal{M}}$ with respect to the exogenous reparametrization $\boldsymbol{\phi} : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{I}}$ given by $\phi_i(\bar{\boldsymbol{e}}) = \Phi^{-1}(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i))$ and for all $i \in \mathcal{I}$

$$\forall \tilde{e}_i \in \tilde{\mathcal{E}}_i, \forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \qquad \tilde{f}_i(\boldsymbol{x}_{\setminus i}, \tilde{e}_i) = \bar{f}_i\big(\boldsymbol{x}_{\setminus i}, (F_{\bar{E}_i}^{\bar{\mathcal{M}}})^{-1}\big(\Phi(\tilde{e}_i)\big)\big).$$

Note that this indeed defines an exogenous reparameterization, as for $\mathbb{P}_{\mathbb{R}^{\mathcal{I}}}$-almost every $\bar{\boldsymbol{e}}$, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$:

$$\tilde{f}_i(\boldsymbol{x}_{\setminus i}, \phi_i(\bar{e}_i)) = \bar{f}_i\big(\boldsymbol{x}_{\setminus i}, (F_{\bar{E}_i}^{\bar{\mathcal{M}}})^{-1}\big(\Phi(\Phi^{-1}(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i)))\big)\big) = \bar{f}_i(\boldsymbol{x}_{\setminus i}, \bar{e}_i) \tag{10}$$

as Φ is invertible and $(F_{\bar{E}_i}^{\bar{\mathcal{M}}})^{-1}\big(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{e}_i)\big) = \bar{e}_i$ $\mathbb{P}_{\mathbb{R}^{\mathcal{I}}}$-a.s. according to Lemma 64.

$\tilde{\mathcal{M}}$ is interventionally equivalent to $\mathcal{M}$, as it is interventionally equivalent to $\bar{\mathcal{M}}$ which is interventionally equivalent to $\mathcal{M}$. Obviously, $\tilde{\mathcal{M}}$ has no self-loops. For all $\boldsymbol{x} \in \mathbb{R}^{\mathcal{I}}$, the function $\tilde{e}_i \mapsto \tilde{f}_i(\boldsymbol{x}_{\setminus i}, \tilde{e}_i)$ is strictly monotonically increasing, as $\bar{e}_i \mapsto \bar{f}_i(\boldsymbol{x}_{\setminus i}, \bar{e}_i)$ is strictly monotonically increasing by assumption and $(F_{\bar{E}_i}^{\bar{\mathcal{M}}})^{-1}$ is strictly monotonically increasing because $F_{\bar{E}_i}^{\bar{\mathcal{M}}}$ is continuous. Finally, note that $\Phi^{-1}(F_{\bar{E}_i}^{\bar{\mathcal{M}}}(\bar{E}_i))$ has a standard-normal distribution as $F_{\bar{E}_i}^{\bar{\mathcal{M}}}$ is continuous.

Equation (9) now follows directly from Lemma 65. □

We are now ready to state the main result of this subsection, which is an example of a real SCM that does not admit a nice reduction. This seems to be due to the nonlinearities in the causal mechanisms in combination with the presence of a confounder, i.e., an exogenous variable with multiple children.

Theorem 69 *There exist real SCMs that do not admit a nice reduction.*

Proof. Consider the following SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{3}, \mathbb{R}^2, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$ with causal mechanisms given by

$$\begin{aligned} x_1 &= f_1(e_1, e_3) = e_1 + e_3 \\ x_2 &= f_2(x_1, e_2, e_3) = x_1 e_3 + e_2 \end{aligned}$$

and where $\mathbb{P}_{\mathbb{R}^3}$ is the standard-normal measure on $\mathbb{R}^3$. See Figure 8 for the corresponding augmented graph. We have

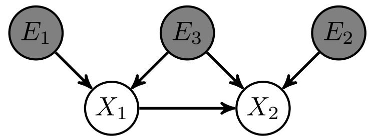

Figure 8: Augmented graph of the SCM in the proof of Theorem 69.

$$F_{X_2}^{\mathcal{M}_{\mathrm{do}(1,\xi)}}(x_2) = \mathbb{P}(\xi E_3 + E_2 \leq x_2) = \Phi\left(\frac{x_2}{\sqrt{1+\xi^2}}\right). \tag{11}$$

If the SCM admits a nice reduction, then it admits one with (marginally) standard-normally distributed exogenous variables according to Lemma 68. Let $(\tilde{f}_1, \tilde{f}_2)$ be the causal mechanisms of that "standard-normal" reduction. According to (9), for all ξ, for $\mathbb{P}_{\mathbb{R}^2}$-almost every $\tilde{\boldsymbol{e}}$:

$$\tilde{f}_2(\xi, \tilde{e}_2) = (F_{X_2}^{\mathcal{M}_{\mathrm{do}(1,\xi)}})^{-1}\big(\Phi(\tilde{e}_2)\big) = \sqrt{1+\xi^2}\,\tilde{e}_2\,,$$

where we used (11) for the second equality.

Let $(\tilde{X}_1, \tilde{X}_2, \tilde{E}_1, \tilde{E}_2)$ be a solution of $\tilde{\mathcal{M}}$. Then $\tilde{X}_2 = \tilde{f}_2(\tilde{X}_1, \tilde{E}_2) = \sqrt{1+\tilde{X}_1^2}\,\tilde{E}_2$ a.s., so

$$\tilde{E}_2 = \frac{\tilde{X}_2}{\sqrt{1+\tilde{X}_1^2}} \text{ a.s..}$$

Now let $(X_1, X_2, E_1, E_2, E_3)$ be any solution of $\mathcal{M}$. By observational equivalence, $(\tilde{X}_1, \tilde{X}_2)$ has the same distribution as (X_1, X_2), and therefore $\tilde{E}_2$ is distributed as

$$\frac{X_2}{\sqrt{1+X_1^2}} = \frac{(E_1+E_3)E_3 + E_2}{\sqrt{1+(E_1+E_3)^2}} \text{ a.s..}$$

This contradicts the fact that $\tilde{E}_2$ must be standard-normally distributed. Indeed, it does not even have expectation value 0:

$$\mathbb{E}\left(\frac{(E_1+E_3)E_3 + E_2}{\sqrt{1+(E_1+E_3)^2}}\right) = \mathbb{E}\left(\frac{(E_1+E_3)E_3}{\sqrt{1+(E_1+E_3)^2}}\right) \neq 0$$

as one can check numerically, for example. □

6 Conclusion and discussion

To the best of our knowledge, this work provides the first rigorous treatment of cyclic SCMs that deals with the measure-theoretic complications of those. We have given formal definitions of SCMs, their solutions, various equivalence classes, their causal interpretation, and we have defined several operations that can be performed on SCMs. Table 1 summarizes some of the preservation properties of these operations.

The central topic of investigation was how to arrive at a parsimonious SCM that describes a subsystem of a system represented by a given SCM.

We decomposed this question into two parts, dealing with the endogenous and exogenous variables separately. For the endogenous variables, we defined a marginalization operation that reduces the endogenous variables in the SCM to a particular subset of interest. For the exogenous variables, we gave several conditions under which a reduction of the number of exogenous variables is possible. One way to obtain a reduction is by using an exogenous reparameterization. We showed that for linear SCMs one can reduce the exogenous variables down to a single one-dimensional exogenous variable per endogenous variable, while preserving linearity. More generally, one can use an exogenous reparameterization with respect to the measure-theoretic isomorphism $\boldsymbol{\phi} : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{I}}$ to get a reduction. However, in general the resulting exogenously reparametrized causal mechanisms will not be very "smooth", and will be hard to estimate from data. We showed that it is not possible in general to find reductions that satisfy certain additional "smoothness" constraints.

At this stage, we can only speculate on the implications of this "negative" result. If such "smooth" reductions would have existed in general, the estimation of an SCM that describes a subsystem of a larger system could be reduced to estimating the causal mechanisms and a probability distribution on a space whose dimensionality equals the number of endogenous variables. The big arsenal of statistical estimation techniques, either parameteric or non-parametric, could then be applied in this context to solve the statistical estimation problem, without the need to make strong assumptions about the number and nature of the latent variables. In the light of our result, however, in order to accurately represent the model, various hypothetical and possibly large exogenous spaces may need to be considered. In the worst case scenario, the number of latent exogenous variables that one needs in order to obtain a consistent estimator may be unlimited. This outcome would contrast with recent findings of Evans and Richardson (2014) for the closely related acyclic directed mixed graphs (ADMGs), who showed that for discrete variables, these models always admit parsimonious parameterizations. Our result may not be the end of the story, though. It is still possible that other types of reductions do exist in general and would be relatively easy to estimate from data. We leave this question for future work.

Regarding the use of measure theory in this work, the reader may wonder whether it was really necessary to invoke some of the more advanced results of measure theory here, while in most treatments of SCMs measure-theoretical aspects are not discussed at all. That this seems necessary indeed is a somewhat unexpected consequence of our decision to develop a general theory that also incorporates cyclic SCMs. Indeed, cycles in SCMs add several technical complications that do not play a role in the acyclic case: unique solvability is no longer guaranteed, nor is it preserved under interventions; solutions may not be uniquely defined, which makes it difficult to define an SCM in terms of random variables; endogenous random variables can no longer be defined "recursively" in terms of each other and the exogenous variables by mere composition of causal mechanisms (which provides a trivial proof of the measurability of the variables in the acyclic case). These technical complications may explain why in most of the SCM literature so far (except possibly for the

Operation	Property	Preserved?
Intervention	acyclicity	True
	uniquely solvable	False
	not uniquely solvable	False
Marginalization	acyclicity	True
	uniquely solvable	True
	not uniquely solvable	True
Exogenous reparametrization	acyclicity	True
	uniquely solvable	True
	not uniquely solvable	True

Table 1: Preservation properties of various operations that can be performed on SCMs.

linear SEM literature), acyclicity has been assumed, even though there is a real need for a theory of cyclic SCMs given that many causal systems in nature involve feedback. Actually, it is surprising that many systems appear to be acyclic at a macroscopic level, even though on a microscopic level, all particles interact with each other, leading to a fully connected cyclic causal graph on that level of detail.

With this work, we hope to have provided a foundation to the theory of cyclic SCMs that will enable such models to be used for the purposes of causal discovery and prediction. Future work consists of developing suitable graphical representations for general SCMs that give rise to Markov properties that allow to read off all generic conditional independences between endogenous variables from the graph, e.g., by generalizing the recent work of Evans (2016) to the cyclic case. This will then enable the development of a novel constraint-based algorithms that can, under suitable assumptions, estimate the structure of SCMs from data. Furthermore, we plan to apply the theory derived here in order to further extend the theory in Mooij et al. (2013) on the close connections between ODEs and SCMs.

Acknowledgments

We thank Patrick Forré and Robin Evans for helpful discussions and Tineke Blom and Alexander Ly for providing helpful comments on earlier drafts. Stephan Bongers and Joris Mooij were supported by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410 and VENI grant 639.031.036). Joris Mooij was also supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 639466).

References

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

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

Dudley, R. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, New York, USA.

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

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

Evans, R. J. and Richardson, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *The Annals Of Statistics*, 42(4):1452–1482.

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

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12.

Hyttinen, A., Eberhardt, F., and Hoyer, P. (2012). Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439.

Kechris, A. (1995). *Classical Descriptive Set Theory*, volume 16 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, USA.

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

Lauritzen, S., Dawid, A., Larsen, B., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, 20:491–505.

Mason, S. (1953). Feedback theory-some properties of signal flow graphs. In *Proceedings of the IRE*, volume 41.

Mason, S. (1956). Feedback theory-further properties of signal flow graphs. In *Proceedings of the IRE*, volume 44.

Mooij, J. M., Janzing, D., and Schölkopf, B. (2013). From ordinary differential equations to structural causal models: the deterministic case. In Nicholson, A. and Smyth, P., editors, *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 440–448. AUAI Press.

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(32):1–102.

Neal, R. (2000). On deducing conditional independence from d-separation in causal graphs with feedback. *Journal of Artificial Intelligence Research*, 12:87–91.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, 1st edition.

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

Pearl, J. and Dechter, R. (1996). Identifying independence in causal graphs with feedback. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 420–426.

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.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, 30(4).

Spirtes, P. (1994). Conditional independence in directed cyclic graphical models for feedback. Technical Report CMU-PHIL-54, Carnegie Mellon University.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 491–499.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search.* The MIT Press, Cambridge, Massachusetts, 2nd edition.

Tian, J. (2002). Studies in causal reasoning and learning. Technical Report R-309, Cognitive Systems Laboratory, University of California, Los Angeles.

Verma, T. (1993). Graphical aspects of causal models. Technical Report R-191, Computer Science Department, University of California, Los Angeles.

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