CAUSAL MODELING & DYNAMICAL SYSTEMS: A NEW PERSPECTIVE ON FEEDBACK

Stephan R. Bongers

CAUSAL MODELING & DYNAMICAL SYSTEMS: A NEW PERSPECTIVE ON FEEDBACK

Stephan R. Bongers

CAUSAL MODELING & DYNAMICAL SYSTEMS: A NEW PERSPECTIVE ON FEEDBACK

Academisch Proefschrift

ter verkrijging van de graad doctor aan
de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een
door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 29 juni 2022, te 14:00 uur

door

Stephan Robert Bongers

geboren te Leusden

PROMOTIECOMMISSIE

Promotor:

prof. dr. J.M. Mooij	Universiteit van Amsterdam

Copromotor:

prof. dr. M. Welling	Universiteit van Amsterdam

Overige leden:

prof. dr. J. Pearl	University of California, Los Angeles
prof. dr. T.S. Richardson	University of Washington
prof. dr. M.R.H. Mandjes	Universiteit van Amsterdam
prof. dr. F.P. Pijpers	Universiteit van Amsterdam
dr. D. Janzing	Amazon Research
dr. S. Magliacane	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Funding for this research was provided by the NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 639466).

Printed by Ridderprint, The Netherlands.

ISBN: 978-94-6458-332-8

SUMMARY

In this thesis, *Causal Modeling & Dynamical Systems: A New Perspective On Feedback*, we propose novel solutions to causal modeling in the presence of latent confounding ("latent common causes") and cyclic causal relationships ("feedback loops"). Our proposed solutions bridge the gap between the worlds of causal models and dynamical systems.

Our main contributions are as follows:

- We propose a **general theory of statistical causal modeling with structural causal models (SCMs)** suitable for modeling latent confounding, cyclic, and nonlinear causal relationships (Bongers et al., 2021; Chapter 2). We show that in the presence of cycles, many convenient properties of acyclic SCMs do not hold in general, such as the existence of a (unique) solution or that of a Markov property. We prove that for SCMs in general, many of these convenient properties hold under certain **solvability conditions**.

- We provide a **marginalization operation** for SCMs (Bongers et al., 2021; Section 2.5), suitable for obtaining a marginal SCM on a subset of the variables. We show that for cyclic SCMs, marginalization does not always exist without further assumptions. We prove that this marginalization operation preserves the probabilistic and causal semantics under certain local unique solvability conditions. Similarly, one can marginalize the graph of an SCM, called the "latent projection" of the graph. We show that, in general, the marginalization of an SCM does not respect the latent projection of its associated graph, but we prove that it does under an additional local ancestral unique solvability condition.

- We provide **conditions for identifying directed paths and bidirected edges** in the graph of an SCM (Bongers et al., 2021; Section 2.7). We show that the presence or absence of a (bi-)directed path or edge cannot always be identified from a difference in observational and/or interventional distributions. Moreover, if cycles are present, "nonancestral" effects may exist, that is, an intervention on a variable may change the distribution of some of its nondescendants in the graph. We prove that this counterintuitive behavior of "nonancestral" effects will not happen under suitable solvability conditions.

- We propose **simple SCMs** (Bongers et al., 2021; Section 2.8). The class of simple SCMs extends the subclass of acyclic SCMs to the cyclic setting while preserving many of their convenient properties, such as the existence and uniqueness of observational and interventional distributions, being closed under intervention and marginalization, satisfying a Markov property. We illustrate that the class of simple SCMs forms a convenient and practical

extension of acyclic SCMs that can be used for causal modeling, learning, and reasoning.

- We propose **structural dynamical causal models (SDCMs)** (Bongers, Blom, and Mooij, 2022; Section 3.3). The SDCM framework enables modeling of stochasticity, time-dependence, and causality in a natural way, and contains the classes of structural causal models (SCMs) and random differential equations (RDEs) as special cases. An SDCM can be thought of as the stochastic-process version of an SCM, where the static random variables of the SCM are replaced by dynamic stochastic processes and their derivatives. We provide a graphical representation for SDCMs and conditions for **the existence and uniqueness of solutions for given initial conditions**. We demonstrate that SDCMs provide the basis for modeling the causal mechanisms that underlie the dynamics of systems encountered in science and engineering.

- We provide an **equilibration operation** for SDCMs (Bongers, Blom, and Mooij, 2022; Section 3.4), suitable for equilibrating an SDCM to an SCM such that the static solutions of the SCM contain the equilibrium solutions of the SDCM, without requiring any assumption on the number of equilibrium solutions of the SDCM. This establishes a bridge between the frameworks of SDCMs and SCMs at equilibrium, which sheds some new light on the causal interpretation of SCMs, particularly on the counterintuitive behavior of "nonancestral" effects at equilibrium. This bridge enables one to study the causal semantics of a large class of stochastic dynamical systems, including those that have **multiple equilibria**.

- We propose a **Markov property** for SDCMs (Bongers, Blom, and Mooij, 2022; Section 3.3.7) that is suitable for both the solutions of the SDCM and the evaluation of the solutions at any point in time under certain conditions.

CONTENTS

1

INTRODUCTION AND BACKGROUND

In this chapter, we introduce and explain background material about probabilistic models, probabilistic graphical models, causal models, and dynamical models. We end the chapter with a short introduction to the research questions studied in this thesis.

1.1 ARTIFICIAL INTELLIGENCE AND CAUSALITY

Human beings are remarkably proficient at identifying relevant objects and concepts that enable them to reason about which actions to take in a given situation. Our species can do this even without an encompassing understanding of all the underlying mechanisms and natural laws at work. Infants, for instance, understand the physical world around them by relying on objects that can be tracked over time in a consistent way (Lake et al., 2017; Dehaene, 2020). At an early age, they can infer that objects act upon each other when they come into contact with each other (Spelke, 1990). This ability of children to build a conceptual representation of the world on which they can perform *causal reasoning*, allows them to quickly learn new tasks, as previously acquired knowledge and understanding of the world can be re-used and re-evaluated. This ability of humans to solve real-world tasks by re-using and re-purposing their knowledge and skills in novel scenarios lies at the heart of our *intelligence*.

In the field of *machine learning*, which is one of the most widely pursued branches in *artificial intelligence* (AI), one of the main objectives is to build machines that themselves can acquire new knowledge and skills through experience in the form of data. A common approach to the problem of learning is to fit a model to data in the hope that this learned model will generalize well to new data or experiences (Mitchell, 1997). Despite its success, this approach to learning provides a rather superficial description of reality that only holds if new data and experiences are coming from a distribution that does not differ (too much) from the training distribution.

The field of *causal learning* seeks to learn a model that fits not only the data at hand but can also describe the effect of intervention in terms of changes in distribution. A causal model differs from a statistical model in that a set of variables determines each variable through a causal relationship, called the *causal mechanism*, that remains *invariant* when other mechanisms are subjected to intervention. This *invariance* means that mechanisms can vary independently of one another, which can and will happen under different experimental conditions (Pearl, 2009). This invariance property of the causal mechanisms allows for a modular representation of the world. Each module represents a causal mechanism, for which some can

behave similarly across different tasks and environments. When learning a causal model, one only needs to adapt a few modules in its internal representation of the world, while others can be re-used without further training. Learning such causal models has been shown to be more efficient and allows for better generalization (see, e.g., Schölkopf et al., 2021).

A core problem in causal learning is that the used models often rest on the *assumption of acyclicity*, or in other words, that no feedback loops are allowed. Although this assumption can be a reasonable approximation and simplifies the theoretical analysis, in many practical cases *feedback loops* are present and should not simply be ignored. Feedback is a very common phenomenon. For example, it is involved in how our body keeps a constant temperature, how the price of a product is determined, and how learning is facilitated throughout our education system. The aim of this work[1] is to further advancement of the field of causal modeling in the presence of feedback.

1.2 PROBABILISTIC MODELS

Machine learning models are often described in terms of *probabilistic models* which provide a statistical description of the system of interest. In probabilistic models, we assume some level of uncertainty over the values of the variables $\boldsymbol{x} := (x_1, \ldots, x_n)$ in the model. A probabilistic model is usually represented in terms of random variables. Let $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}$ be the observed random variable, that assigns a probability to each (measurable) subset of $\boldsymbol{\mathcal{X}}$, where Ω denotes the *sample space* of some background probability space[2] $(\Omega, \mathcal{F}, \mathbb{P})$ and $\boldsymbol{\mathcal{X}}$ is some space[3] of interest. For simplicity, we assume in this chapter that random variables have a probability density,[4] which for $\boldsymbol{X}$ is given by the (joint) probability density

$$p(\boldsymbol{x}) := p(\boldsymbol{X} = \boldsymbol{x}) .$$

We will use capital letters (e.g., $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) for random variables and lowercase letters ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$) for values taken by the corresponding random variables. The statement $\boldsymbol{X} = \boldsymbol{x}$ describes an *event* which corresponds with a set of samples in the sample space Ω.

1 Our work is strongly influenced by Pearl's account on the mathematization of causal modeling (Pearl, 2009).

2 A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ consists of a measurable space $(\Omega, \mathcal{F})$, where Ω is a non-empty set and $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra, together with a probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ which is a finite measure on $\mathcal{F}$ (i.e., σ-additive) such that $\mathbb{P}(\Omega) = 1$.

3 We assume here that $\boldsymbol{\mathcal{X}}$ is a *standard measurable space* (see Definition 2.F.1 in Appendix 2.F) for which a general existence result for regular conditional probability distributions holds (see, e.g., Klenke, 2014).

4 More precisely, we assume that they have a probability density w.r.t. some product measure. For a random variable $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}$, a measure $\boldsymbol{\mu}$ on the measurable space $\boldsymbol{\mathcal{X}}$, and a measurable map $p : \boldsymbol{\mathcal{X}} \to [0, \infty)$, we say that a *random variable $\boldsymbol{X}$ has a probability density $p(\boldsymbol{x})$ w.r.t. a measure $\boldsymbol{\mu}$*, if for all measurable sets U of $\boldsymbol{\mathcal{X}}$ we have

$$\mathbb{P}^{\boldsymbol{X}}(U) = \int_U p(\boldsymbol{x}) d\boldsymbol{\mu} ,$$

where $\mathbb{P}^{\boldsymbol{X}}$ is the induced probability distribution of $\boldsymbol{X}$ on $\boldsymbol{\mathcal{X}}$.

Consider a probabilistic model with joint probability density $p(\boldsymbol{x}, \boldsymbol{y})$. There are two basic operations for determining the probabilities of events of interest, namely the marginalization and conditioning operation. The *marginalization operation* corresponds to summing up/integrating probabilities over all values of a specific variable $\boldsymbol{Y}$, also called *marginalization* over $\boldsymbol{Y}$, and gives the *marginal probability density*

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \,.$$

This identity which connects the joint to the marginal probability density, is also known as the *sum rule* and allows us to look only at a subsystem of variables of interest. The *conditioning operation* specifies the probability in $\boldsymbol{y}$ given that $\boldsymbol{x}$ is known with absolute certainty, and gives the *conditional probability density*

$$p(\boldsymbol{y} \mid \boldsymbol{x})$$

which satisfies

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}) p(\boldsymbol{y} \mid \boldsymbol{x}) \,.$$

This identity is also known as the *product/chain rule* and results in the well-known *Bayes' rule*

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{x}) p(\boldsymbol{y} \mid \boldsymbol{x})}{p(\boldsymbol{y})} \,.$$

Often in learning problems, such as in classification or regression problems, we are not interested in the probability density $p(\boldsymbol{x})$, but rather in the conditional probability density $p(\boldsymbol{y} \mid \boldsymbol{x})$, in which case $\boldsymbol{x}$ and $\boldsymbol{y}$ denote respectively the input and target of the model.

Conditional models can be insensitive to certain input variables, which is captured by the notion of conditional independence. Two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be *conditionally independent given* $\boldsymbol{Z}$, denoted by

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \,,$$

if

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}) \,,$$

whenever $p(\boldsymbol{y}, \boldsymbol{z}) > 0$. In other words, knowing the value of $\boldsymbol{Y}$ does not provide any additional information about $\boldsymbol{X}$ given that we know $\boldsymbol{Z}$.

1.3 PROBABILISTIC GRAPHICAL MODELS

A *probabilistic graphical model* is a probabilistic model for which a graph expresses the conditional independencies between random variables. The graph provides an economical representation of the probabilistic model, allowing for efficient inference and expressing different modeling assumptions (i.e., causal modeling assumptions).

1.3.1 *Bayesian networks*

One of the most commonly known probabilistic graphical models are *Bayesian networks*, also known as *probabilistic graphical models for directed acyclic graphs (DAGs)*. A DAG consists of nodes linked by directed arrows such that there is no cycle in the graph. The joint probability density over the variables of such a probabilistic model satisfies the *recursive factorization property relative to a DAG* $\mathcal{G}$, that is,

$$p(x_1, \ldots, x_n) = \prod_i^n p(x_i \mid \boldsymbol{x}_{\mathrm{pa}_{\mathcal{G}}(i)}) \,,$$

where $\mathrm{pa}_{\mathcal{G}}(i)$ denotes the parents of node i in the DAG $\mathcal{G}$. For non-root nodes, a factor corresponds to a conditional probability density, where we condition on its parents. For root nodes, the set of parents is the empty set such that the probability density is unconditional.

The elegance of *Bayesian network* is rooted in the equivalence of the recursive factorization property and various versions of the *Markov property* (Lauritzen et al., 1990; Lauritzen, 1996; Forré and Mooij, 2017). The directed global Markov property for DAGs, also known as the *d-separation criterion* (Pearl, 1985), is one of the most widely used Markov properties. A probabilistic model satisfies the *directed global Markov property relative to a DAG* $\mathcal{G}$ (see Definition 2.A.6 in the Appendix 2.A.2 for a more exact formulation) if for all subsets A, B and C of nodes in $\mathcal{G}$

$$A \overset{d}{\underset{\mathcal{G}}{\perp}} B \,|\, C \quad \Longrightarrow \quad \boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \,|\, \boldsymbol{X}_C \,,$$

where the term on the left reads as A is d-separated from B given C in $\mathcal{G}$ (see Definition 2.A.4). This allows us to read off a set of conditional independence relations that are satisfied in the probabilistic model from the graph by checking which d-separation statements hold. We can see the graph of the model as a carrier of independence assumptions.

Probabilistic graphical models for DAGs have two major shortcomings, namely that (i) they are not closed under marginalization/latent projection,[5] and (ii) they do not allow for cycles in the directed graph. This hinders application where we almost always have incomplete data and/or some feedback loops between observed variables. Several extensions have been proposed for these models to address the problem of not being closed under marginalization/latent projection, resulting in probabilistic graphical models for *acyclic directed mixed graphs* (ADMGs) (Richardson, 2003), i.e., DAGs with bidirected edges, and more generally, for *marginalized directed acyclic graphs* (mDAGs) (Evans, 2016), i.e., DAGs with hyperedges. The problem of not allowing for cycles has been addressed in the case of discrete (Pearl and Dechter, 1996; Neal, 2000) and linear models (Spirtes, 1993, 1994, 1995; Koster, 1996; Richardson, 1996c; Hyttinen, Eberhardt, and Hoyer, 2012). The more general case where nonlinear relationships are allowed was recently addressed by (Forré and

5 The *latent projection* is a marginalization operation for directed graphs (see Definition 2.5.7 and Verma (1993)).

Mooij, 2017), where they addressed both problems (i) and (ii) at once, resulting in general probabilistic graphical models for *HEDGes*, i.e., directed graphs with hyperedges.

1.4 CAUSAL MODELS

We use the term *causal model* to denote a model that describes the *causal mechanisms* of a system. The causal mechanisms of such a model are assumed to stay *invariant* when other mechanisms are subjected to intervention. This *modularity* assumption allows us to describe the effect of external intervention by only changing the affected causal mechanisms/modules while letting the others remain invariant.

1.4.1 *Causal Bayesian networks*

A well-known class of causal models are *causal Bayesian networks* (Pearl, 2009), which are Bayesian networks where each factor $p(x_i \mid \boldsymbol{x}_{\mathrm{pa}_{\mathcal{G}}(i)})$ in the joint probability density represents a causal mechanism of the variable X_i. Perhaps the simplest causal Bayesian network is one that is specified by the following recursive factorization

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}) p(\boldsymbol{y} \mid \boldsymbol{x}) \,,$$

which graph in terms of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is depicted in Figure 1.1 (left). The directed edge in the graph represents a causal relationship, namely that $\boldsymbol{X}$ is a cause of $\boldsymbol{Y}$. The causal mechanism $p(\boldsymbol{y} \mid \boldsymbol{x})$ determines the probability of $\boldsymbol{Y}$ for every intervened value $\boldsymbol{X} = \boldsymbol{x}$. The causal mechanism $p(\boldsymbol{x})$ of $\boldsymbol{X}$ does not change after any intervention on $\boldsymbol{Y}$, since it is an unconditional probability density. This illustrates that causation is an inherently asymmetric concept, manipulating a cause will change its effect, not necessarily vice versa. This shows that the graph of a causal Bayesian network is not only a carrier of independence assumptions but also a carrier of causal assumptions.

The close connection between causality and statistical dependence was already postulated by Reichenbach (1956), who postulated the *common cause principle*,[6] also known as "no correlation without causation". Informally, this principle states that, if two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are statistically dependent, then either one is a cause of the other or they have a common cause that renders them conditionally independent. This postulate can be derived from the framework of causal Bayesian networks by observing that a statistical dependence between variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be described by any of the causal Bayesian networks depicted in Figure 1.1. This means that any observational distribution over $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be realized by any of the three models. Thus, without any causal assumptions, the three models cannot be distinguished from each other. Therefore, a causal model can be more informative than a probabilistic one.

6 This principle implicitly assumes no *selection bias*, i.e., there is no conditioning on a (latent) common effect.

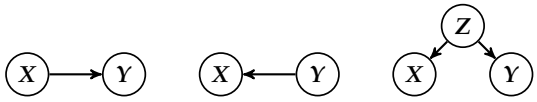
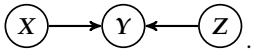


Figure 1.1: Graphs of different causal Bayesian networks with different recursive factorizations.

The task of causal learning goes beyond learning a model based on statistical associations alone because these models allow us to exploit causal assumptions and causal knowledge. But how do we acquire causal knowledge in the first place? The gold standard for *causal discovery* is *randomized controlled experimentation* (Fisher, 1935). The basic idea in such experiments is to randomly assign subjects to a treatment and control group from which we can estimate the effect of the treatment/intervention by comparing the difference in outcome between both groups. This forms the basis of many scientific experiments; however, it can often be difficult, unethical, or too expensive to perform such an experiment. To remedy this, researchers found ways to acquire causal knowledge, not from randomized controlled experimentation but from purely observational data instead. The key observation was that one could recover causal relationships from the statistical patterns in the data. For example, Rebane and Pearl (1987) observed that one could determine the causal direction between two variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ by observing that $\boldsymbol{X}$ correlates with $\boldsymbol{Y}$ and that there exists a third variable $\boldsymbol{Z}$ that correlates with $\boldsymbol{Y}$ but not with $\boldsymbol{X}$, as in the *collider*



.

This led to various causal discovery methods, such as the IC and PC algorithm for causal Bayesian networks (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009).

The simplicity of causal Bayesian networks, by interpreting the conditional factors as causal mechanisms, makes them conceptually appealing. However, they still inherit the shortcomings (i) and (ii) of Bayesian networks. The problem that causal Bayesian networks are not closed under marginalization/latent projection and do not allow for cycles is rooted in the recursive factorization property. A better way to define the causal mechanisms is in terms of functional parent-child relationships rather than conditional factors, which provide a new perspective on causality.

1.4.2 *Structural causal models*

In this work, we focus on the class of *structural causal models* (SCMs) (Pearl, 2009). Traditionally, in a structural causal model[7] the causal mechanism of a variable X_i is described by the assignment

$$X_i = f_i(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{G}}(i)}, E_i)\,,$$

7 Formal treatment of structural causal models with cycles and latent variables is given in Chapter 2.

where f_i is a deterministic measurable function depending on the parents of X_i in the directed graph $\mathcal{G}$ and E_i an unobserved random variable that is not in $\mathcal{G}$. We call the observed random variables X_i and unobserved random variables E_i respectively the *endogenous* and *exogenous* variables. Often, the exogenous variable E_i ensures us that endogenous variable $X_i = f_i(\boldsymbol{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, E_i)$ represents a conditional probability $p(x_i \mid \boldsymbol{x}_{\mathrm{pa}_{\mathcal{G}}(i)})$. The assignments are also called the *structural equations* of the model. The structural equations are assumed to stay *invariant* when other equations are subjected to intervention. Interventions can be straightforwardly formalized as an operation that modifies a subset of the structural equations and the graph accordingly. The simplest type of intervention is the *perfect intervention* which forces a certain variable, say X_i, to take on some fixed value x_i. Such a perfect intervention, denoted by $\mathrm{do}(X_i = x_i)$, replaces the structural equations of X_i by the intervened structural equation $X_i = x_i$, whereby we remove all edges in the graph $\mathcal{G}$ that have an arrowhead pointing towards X_i.

A special subclass of SCMs is the class of *acyclic SCMs*, where the directed graph $\mathcal{G}$ is acyclic. If, in addition, the random variables $E_1, \ldots, E_n$ are jointly independent, then the model is called a *Markovian SCM*. For a Markovian SCM, various equivalent versions of the Markov property and the equivalent recursive factorization property hold, and gives the same recursive factorization that characterizes causal Bayesian networks. Acyclic SCMs, also known as *semi-Markovian SCMs*, have been widely studied and are well-understood (see, e.g., Verma, 1993; Richardson, 2003; Evans, 2016). Although they do not satisfy a recursive factorization property that factors all the causal mechanisms, they do obey various equivalent versions of the Markov property. Furthermore, they induce a unique distribution over the observed variables, and they are closed under marginalization/latent projection.

An important advantage of SCMs over causal Bayesian networks is that they can represent cyclic causal relationships. To give a concrete example, consider the linear model given by the structural equations

$$\begin{aligned} X_1 &= E_1 \\ X_2 &= E_2 \\ X_3 &= \alpha X_4 + X_1 + E_3 \\ X_4 &= \beta X_3 + X_2 + E_4\,, \end{aligned}$$

where $\alpha, \beta \neq 0$, $\alpha\beta \neq 1$, and $E_1, \ldots, E_4$ are jointly independent Gaussian random variables. Its graph is depicted in figure 1.2. Performing different perfect interventions on X_3 yields different probabilities for X_4, and vice versa. Although there does not exist a recursive factorization property for this model, the global directed Markov property holds for this model and for linear SCMs in general (Spirtes, 1993, 1995; Forré and Mooij, 2017).

Although some progress has been made in the discrete (Pearl and Dechter, 1996; Neal, 2000) and linear case (Spirtes, 1994, 1995; Koster, 1996; Richardson, 1996c; Hyttinen, Eberhardt, and Hoyer, 2012), one encounters various technical complications in the general cyclic case. For example, Spirtes (1994, 1995) showed

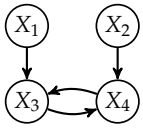


Figure 1.2: Graph of a cyclic SCM.

that the directed global Markov property in terms of the d-separation criterion does not hold anymore for the slightly modified nonlinear model (see Figure 1.2)

$$\begin{aligned} X_1 &= E_1 \\ X_2 &= E_2 \\ X_3 &= X_1 X_4 + E_3 \\ X_4 &= X_2 X_3 + E_4 \,, \end{aligned}$$

and proposed an alternative criterion in terms of a "collapsed graph". More recently, Forré and Mooij (2017) showed that an alternative formulation of the Markov property in terms of the *σ-separation criterion*, a generalized version of the d-separation criterion, holds for such general cyclic probabilistic graphical models.

The main difficulty of such SCMs with cycles is that many convenient properties do not hold anymore after intervention. For example, performing the intervention $\mathrm{do}(X_1 = -1, X_2 = -1)$ on the previous model gives an intervened model for which σ-separation does not hold anymore. Furthermore, the structural equations have either multiple solutions for the random variables X_3 and X_4 or no solution at all (depending on E_3 being equal to E_4 or not), which hinders the causal interpretation of such models.

1.5 DYNAMICAL SYSTEMS

The task of causal modeling is to construct a modular representation of a system in terms of causal mechanisms. However, many systems in the real world are represented by dynamical systems described by a system of differential equations that allow for modeling time-dependent behavior. An *ordinary differential equation* is a system of coupled differential equations which relate the time derivative of each function $X_i(t)$ to the functions $\boldsymbol{X}(t)$ as follows

$$\frac{dX_i}{dt} = f_i(\boldsymbol{X}) \,,$$

where f_i is a deterministic function depending on $\boldsymbol{X}$ and $\boldsymbol{X}$ is a function on $\mathbb{R}$ that takes values in $\mathbb{R}^n$. An ordinary differential equation together with an *initial value* $\boldsymbol{X}(t_0) = \boldsymbol{X}_{[0]} \in \mathbb{R}^n$ at $t_0 \in \mathbb{R}$ is called an *initial value problem*. The Picard-Lindelöf theorem guarantees that, at least locally, if $\boldsymbol{f}$ is Lipschitz, then the initial value problem has a unique solution $\boldsymbol{X}(t)$ for $t \geq t_0$ and is of the form

$$\boldsymbol{X}(t) = \boldsymbol{X}_{[0]} + \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{X}) dt \,.$$

Thus, $\boldsymbol{X}(t)$ is, at least locally, deterministically determined by its past values. Such *temporal precedence* can be interpreted as that "the cause always precedes its effect in time", which is often assumed to be essential for defining causation. However, temporal information often is not enough to distinguish causal effects from spurious associations caused by unknown factors (Pearl, 2009). For example, that the rooster crows immediately before sunrise does not imply that the rooster causes the sun to rise. Assuming that one thing preceding another can be used as a proof of causation without taking all the unknown factors into account is also known as the "post hoc ergo propter hoc" fallacy[8] and may lead to false causal knowledge.

Over the years, several efforts have been made to develop a notion of causality for dynamical systems. One approach is to discretize time by intervals Δt such that the differential equation can be rewritten as a difference equation given by

$$\boldsymbol{X}(t + \Delta t) = \boldsymbol{X}(t) + \Delta t \cdot \boldsymbol{f}(\boldsymbol{X}(t)) \,.$$

For any point in time, the future of $\boldsymbol{X}(t)$ is determined by this equation. Discretizing time has the benefit that one can easily incorporate stochasticity in the model by taking $\boldsymbol{X}(t)$ a random variable at every point in time. Often, one allows for additional stochasticity in the parameters in the model by making $\boldsymbol{f}$ dependent on an additional random noise function $\boldsymbol{E}(t)$ that, for example, can model different unknown factors. Examples of such models are the simultaneous equation models (Fisher, 1970; Lacerda et al., 2008), vector autoregressive (VAR) models (Sims, 1980; Lütkepohl, 2005) and dynamic Bayesian networks (Dagum, Galper, and Horvitz, 1992; Ghahramani, 1998). Often, a causal interpretation is given to such models by assuming that the causal mechanisms are correctly described by the function $\boldsymbol{f}$, and possibly $\boldsymbol{E}$ (assuming that one has taken all unknown factors into account). In principle, these models fit directly into the framework of acyclic SCMs by labeling the random variables with time. For continuous time, a causal interpretation has been given to a system of differential equations that are driven by a certain type of noise, the so-called stochastic differential equations (see, e.g., Florens and Fougere, 1996; Hansen and Sokol, 2014). In the deterministic setting, several approaches attribute a causal interpretation to first-order systems of ODEs (see, e.g., Iwasaki and Simon, 1994; Mooij, Janzing, and Schölkopf, 2013; Pfister, Bauer, and Peters, 2019; Blom and Mooij, 2021). Often, one simply takes f_i as the causal mechanism that describes the equation of motion of each variable X_i. Iwasaki and Simon (1994) study the notion of causality in ODEs using Simon's causal ordering algorithm.

One important advantage of dynamical systems is that they can model feedback. For example, consider the mass-spring system in physics, depicted in Figure 1.3 (top left), that consists of three point masses m_i with positions $Q_i(t)$, between the walls at $Q_0(t) = 0$ and $Q_4(t) = L$ and momenta $P_i(t)$ $(i = 1, 2, 3)$ that are coupled

8 The Latin phrase "post hoc ergo propter hoc" means "after this, therefore because of this". This fallacy informally states that "If an event followed another event, then the first event must have been caused by the second event.".

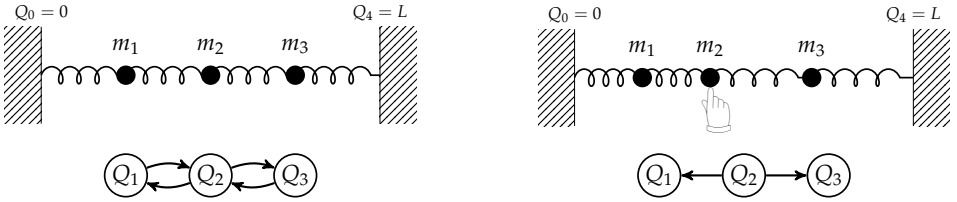


Figure 1.3: Mass-spring system of three point masses m_1, m_2 and m_3 before (top left) and after holding the mass m_2 fixed (top right). The graphs of the SCM that describe the positions of the masses at equilibrium before (bottom left) and after holding m_2 fixed (bottom right).

to each other by springs with spring constant κ_i and length ℓ_i ($i = 0, \ldots, 3$). The equation of motion for each mass m_i can be described by the differential equations

$$\frac{dP_i}{dt} = \frac{k_i}{m_i}(Q_{i+1} - Q_i - \ell_i) + \frac{k_{i-1}}{m_i}(Q_{i-1} - Q_i + \ell_{i-1}) - \frac{b_i}{m_i}P_i$$
$$\frac{dQ_i}{dt} = P_i/m_i\,,$$

where $b_i \geq 0$ ($i = 1, 2, 3$) denotes some friction coefficients. Any two masses connected (i.e., coupled) by a spring may exert forces on each other. Under friction (i.e., $b_i > 0$ for all i), there is a unique equilibrium position where the sum of forces vanishes for every mass m_i. Temporarily moving some masses out of equilibrium position will bring them in a damped oscillatory motion that will converge to their unique stable equilibrium positions.

The stationary behavior of dynamical systems can be described by an SCM under certain stability conditions. In the deterministic setting, Mooij, Janzing, and Schölkopf (2013) showed that for globally asymptotically stable systems, an SCM could be obtained by *equilibrating* a system of first-order ordinary differential equations. The idea is that one obtains a set of *labeled equilibrium equations* by setting the time derivatives of the variables equal to zero and labeling each equation with the variable of its corresponding derivative. If possible, the structural equations of an SCM can be derived from the labeled equilibrium equations by solving them w.r.t. their corresponding variables. For example, the above mass-spring system yields the SCM with structural equations

$$Q_i = \frac{k_i(Q_{i+1} - \ell_i) + k_{i-1}(Q_{i-1} + \ell_{i-1})}{k_i + k_{i-1}}$$
$$P_i = 0$$

for each Q_i and P_i respectively, which graph for the Q_i's is depicted in Figure 1.3 (bottom left). This SCM describes how the equilibrium states of the dynamical systems change under interventions. For example, holding the mass m_2 fixed

at position $\frac{2}{5}L$ by performing the intervention[9] $\mathrm{do}(Q_2 = \frac{2}{5}L, P_2 = 0)$ yields the intervened SCM depicted in Figure 1.3 (right), which illustrates that holding Q_2 fixed at a certain position can have an effect on both Q_1 and Q_3 at equilibrium.

While a system of first-order ODEs provides a rather comprehensive description of a system, learning such a system requires time series data with a sufficiently high temporal resolution, which can be costly, impractical, or even impossible. In the case of a globally asymptotically stable system, the more compact SCM representation, where we discard all the temporal information, has advantages for learning and prediction purposes at equilibrium. On the other hand, first-order ODEs form a well-established and well-understood framework where the modeling of feedback forms no problem at all.

1.6 RESEARCH QUESTIONS AND CONTRIBUTIONS

The contributions of this thesis are guided by the following research questions:

Research question 1: *How can we extend the class of acyclic SCMs to the cyclic setting while preserving many of their convenient properties?*

In chapter 2 and (Bongers et al., 2021) we propose a *general theory of statistical causal modeling with SCMs* suitable for modeling latent confounding, cyclic and nonlinear causal relationships. We provide a definition of an SCM that slightly deviates from previous notions of (acyclic) SCMs because we separate the model from the (endogenous) random variables that solve it. This slight modification makes interventions on SCMs always well-defined, even if the resulting intervened SCM does not have a (unique) solution.

We provide the notions of *(local) solvability* and *unique solvability* (Section 2.3) that describe the existence and uniqueness of (local) measurable solution functions for a subsystem of structural equations. These notions play a central role in extending many convenient properties of acyclic SCM to the cyclic setting. We show that solvability of an SCM is a sufficient and necessary condition for the existence of a solution of an SCM. Moreover, unique solvability implies the uniqueness of the induced observational distribution. We show that, in general, cyclic SCMs may have no solution, solutions with a unique distribution, or solutions with different distributions.

We provide a *marginalization operation* for SCMs (Section 2.5), suitable for obtaining a marginal SCM on a subset of the variables. We show that for cyclic SCMs, marginalization does not always exist without further assumptions. We prove that this marginalization operation preserves the probabilistic and causal semantics under certain local unique solvability conditions. Similarly, one can marginalize the graph of an SCM, which is called the "latent projection" of the graph. We show that, in general, the marginalization of an SCM does not respect the latent projection of

9 Mooij, Janzing, and Schölkopf (2013) only consider interventions for the mass-spring system of the form $\mathrm{do}(Q_i = \xi_i, P_i = 0)$ with $\xi_i \in \mathbb{R}$, because such interventions with nonzero values for the momenta P_i are physically impossible.

its associated graph, but we prove that it does so under an additional local ancestral unique solvability condition.

We provide an overview of *Markov properties* for SCMs (Section 2.6) suitable for reading off conditional independencies in a distribution directly from the graph of the SCM. Markov properties have been of key importance to derive various central results regarding causal reasoning and causal learning. Although the usual Markov properties do not hold in general for cyclic SCMs, we provide an overview under which conditions they hold.

We propose *simple SCMs* (Section 2.8). The class of simple SCMs extends the subclass of acyclic SCMs to the cyclic setting while preserving many of their convenient properties, such as the existence and uniqueness of observational and interventional distributions, being closed under marginalization, satisfying a Markov property. We illustrate that the class of simple SCMs forms a convenient and practical extension of the class of acyclic SCMs that can be used for causal modeling, learning, and reasoning.

As explained in Section 2.3, cyclic SCMs may contain no solutions or (multiple) solutions with either unique or different distributions. This leads us to the following question:

Research question 2: *Can the (multiple) equilibrium states of a dynamical system be described by an SCM?*

In chapter 3 and (Bongers, Blom, and Mooij, 2022) we propose *structural dynamical causal models* (SDCMs), a class of models that can describe a large class of continuous dynamical systems by random differential equations of arbitrary order (including zeroth-order). We provide an *equilibration operation* (Section 3.4.3) suitable for equilibrating an SDCM to an SCM such that the static solutions of the SCM contain all the equilibrium solutions of the SDCM, without requiring any assumption on the number of equilibrium solutions of the SDCM. For example, we demonstrate that in the mass-spring system depicted in Figure 1.3 we can relax the nonzero friction condition in the model, i.e., the system does not have to be globally asymptotically stable.

The proposed framework of SDCMs *enables modeling of arbitrary order differential equations, including zeroth-order equations*. We demonstrate that SDCMs are capable of modeling systems like the well-known price, supply and demand model from economics (see, e.g., Richardson and Robins, 2014) given by the equations

$$\begin{aligned} \frac{dX_P}{dt} &= \lambda(X_D - X_S) \\ X_S &= \beta_S X_P + E_S \\ X_D &= \beta_D X_P + E_D \,, \end{aligned}$$

where X_P, X_S, and X_D denote the price, supply and demand of a quantity of a product respectively, E_S and E_D are some random exogenous influences on the supply and demand respectively, and β_S, β_D, and λ are some constant parameters.

We show that the "market equilibrium" of this system can be modeled by a cyclic SCM. However, allowing for zeroth-order dynamic structural equations leads to additional technical challenges that are absent when solving first-order RDEs. For example, the initial conditions of the solutions may be constrained by the zeroth-order dynamic structural equations, and possibly even by additional "hidden" constraints. In Section 3.3.6, we provide sufficient *conditions under which the existence and uniqueness of a solution* of an SDCM with a given initial condition can be guaranteed.

An additional advantage of the proposed framework of SDCMs is that it *enables the modeling of stochasticity* of the dynamical system by combining random differential equations with additional zeroth-order equations. Random differential equations (RDEs) are similar to ordinary differential equations (ODEs) but can deal with randomness in the initial conditions and the parameters. This enables one to model randomness in the equilibrium solutions and apply statistical tools available for SCMs when studying the equilibrium solutions of stochastic dynamical systems. For example, one can apply the d-separation criterion to the graph of the equilibrated SCM of the intervened mass-spring system depicted in Figure 1.3 (right) to show that $Q_1 \perp\!\!\!\perp Q_3 \mid Q_2$ at equilibrium.

Research question 3: *Does there exist a general causal modeling class for dynamical systems?*

In chapter 3 and (Bongers, Blom, and Mooij, 2022) we demonstrate how the framework of SDCMs can model stochasticity, time-dependence, and causality in a natural way. An *SDCM can be thought of as the stochastic-process version of an SCM* (Section 3.3.2), where the static random variables of the SCM are replaced by dynamic stochastic processes and their derivatives. We consider stochastic interventions to express their causal semantics (Section 3.3.3) and provide a graphical representation representing their model structure (Section 3.3.4). We demonstrate that by no longer restricting to first-order dynamical systems, we arrive at a more natural causal interpretation of systems of higher-order RDEs where we refrain from modeling higher-order derivatives as separate processes, like in the mass-spring system. Thereby, we circumvent questions like "does position cause velocity, or does velocity cause position, or both?". We demonstrate that SDCMs provide the basis for modeling the causal mechanisms that underlie the dynamics of systems encountered in science and engineering.

We show that the *equilibration operation commutes with intervention* and naturally maps the graph of the SDCM to the graph of the SCM (Section 3.4.4 and 3.4.5) without requiring the assumption that all the solutions equilibrate to the same static equilibrium state. In other words, the proposed equilibration operation maps an SDCM to an SCM while preserving the causal semantics. We demonstrate that the inverse mapping holds trivially. In the non-trivial case, we show that for a certain class of SCMs, one can construct a first-order SDCM with non-trivial dynamics for which all the solutions equilibrate to solutions of the SCM independently of the initial conditions, even after intervention. This shows that under certain

conditions, one can *construct a stable SDCM that realizes the causal semantics of the SCM at equilibrium* (Section 3.4.6).

One important advantage of this bridge between the framework of SDCMs and SCMs at equilibrium is that it enables one to leverage the wealth of statistical tools and discovery methods available for SCMs when studying the causal semantics of a large class of stochastic dynamical systems, including those that have multiple equilibria.

Research question 4: *Does the proposed SDCM framework have a Markov property?*

In Section 3.3.7 and (Bongers, Blom, and Mooij, 2022) we propose a *Markov property for SDCMs*, in analogy with that of SCMs that is suitable for both the solutions of the SDCM and the evaluation of the solutions at any point in time under certain conditions. Key to proving this Markov property are the conditions under which the existence and uniqueness of a solution of SDCM can be guaranteed (Section 3.3.6).

Research question 5: *How can we causally interpret the graph in the presence of cycles?*

In Section 2.7 and (Bongers et al., 2021) we provide *conditions for identifying directed paths and bidirected edges* in the graph of an SCM. We show that, in general, the presence or absence of a (bi-)directed path or edge cannot always be identified from a difference in observational and/or interventional distributions. Moreover, if cycles are present, counterintuitive "nonancestral" effects may exist, that is, an intervention on a variable may change the distribution of some of its nondescendants in the graph. We prove this counterintuitive behavior of "nonancestral" effects will not happen under suitable solvability conditions.

In Section 3.4.7, we demonstrate that the counterintuitive behavior of "nonancestral" effects in the equilibrated SCM can be explained by the dependence of the equilibrium states on different initial conditions. We show that one can view this as selection bias due to equilibration. This sheds some new light on the causal interpretation of SCMs and provides a new perspective on feedback systems at equilibrium.

2

STRUCTURAL CAUSAL MODELS WITH CYCLES AND LATENT VARIABLES

Structural causal models (SCMs), also known as (nonparametric) structural equation models (SEMs), are widely used for causal modeling purposes. In particular, acyclic SCMs, also known as recursive SEMs, form a well-studied subclass of SCMs that generalize causal Bayesian networks to allow for latent confounders. In this chapter, we investigate SCMs in a more general setting, allowing for the presence of both latent confounders and cycles. We show that in the presence of cycles, many of the convenient properties of acyclic SCMs do not hold in general: they do not always have a solution; they do not always induce unique observational, interventional and counterfactual distributions; a marginalization does not always exist, and if it exists the marginal model does not always respect the latent projection; they do not always satisfy a Markov property; and their graphs are not always consistent with their causal semantics. We prove that for SCMs in general each of these properties does hold under certain solvability conditions. Our work generalizes results for SCMs with cycles that were only known for certain special cases so far. We introduce the class of simple SCMs that extends the class of acyclic SCMs to the cyclic setting, while preserving many of the convenient properties of acyclic SCMs. In this chapter we aim to provide the foundations for a general theory of statistical causal modeling with SCMs.[1]

2.1 INTRODUCTION

Structural causal models (SCMs), also known as (nonparametric) structural equation models (SEMs), are widely used for causal modeling purposes (Bollen, 1989; Spirtes, Glymour, and Scheines, 2000; Pearl, 2009; Peters, Janzing, and Schölkopf, 2017). They form the basis for many statistical methods that aim at inferring knowledge of the underlying causal structure from data (see, e.g., Maathuis et al., 2009; Mooij and Heskes, 2013; Bühlmann, Peters, and Ernest, 2014; Peters et al., 2014; Mooij et al., 2016). In these models, the causal relationships between the variables are expressed in the form of deterministic, functional relationships, and probabilities are introduced through the assumption that certain variables are exogenous latent random variables. SCMs arose out of certain causal models that were first introduced in genetics (Wright, 1921), econometrics (Haavelmo, 1943), electrical engineering (Mason, 1953, 1956) and the social sciences (Goldberger and Duncan, 1973; Duncan, 1975).

1 This chapter is adapted from our publication (Bongers et al., 2021). Permission was given by the co-authors for reproduction in this thesis.

Acyclic SCMs, also known as recursive SEMs, form a special well-studied subclass of SCMs that generalize causal Bayesian networks (Pearl, 2009). They have many convenient properties (see, e.g., Pearl, 1985; Lauritzen et al., 1990; Verma, 1993; Lauritzen, 1996; Richardson, 2003; Evans, 2016; Evans, 2018): (i) they induce a unique distribution over the variables; (ii) they are closed under perfect interventions; (iii) they are closed under marginalizations; (iv) their marginalization respects the latent projection; (v) they obey (various equivalent versions of) the Markov property and (vi) their graphs express the causal relationships encoded by the SCM in an intuitive manner.

One important limitation of acyclic SCMs is that they cannot model systems that involve causal cycles. In many systems occurring in the real world, there are feedback loops between observed variables. For example, in economics the price of a product may be a function of the demanded or supplied quantities, and vice versa, the demanded and supplied quantities may be functions of the price. The underlying dynamic processes describing such systems have an acyclic causal structure over time. However, causal cycles may arise when one approximates such systems over time (Fisher, 1970; Mogensen, Malinsky, and Hansen, 2018; Mogensen and Hansen, 2020) or when one describes the equilibrium states of these systems (Iwasaki and Simon, 1994; Lacerda et al., 2008; Hyttinen, Eberhardt, and Hoyer, 2012; Mooij, Janzing, and Schölkopf, 2013; Blom, Bongers, and Mooij, 2019; Pfister, Bauer, and Peters, 2019; Bongers, Blom, and Mooij, 2022). In particular, we show in Chapter 3 that the equilibrium states of a system governed by (random) differential equations can be described by an SCM that represents their causal semantics, which gives rise to a plethora of SCMs that include cycles (we provide already some examples of such feedback systems in Appendix 2.D.1). In contrast to their acyclic counterparts, SCMs with cycles have enjoyed less attention in the literature and are not as well understood. In general, none of the above properties (i)–(vi) hold in the class of SCMs. However, some progress has been made in the case of discrete (Pearl and Dechter, 1996; Neal, 2000) and linear models (Spirtes, 1993, 1994, 1995; Koster, 1996; Richardson, 1996c; Hyttinen, Eberhardt, and Hoyer, 2012), and more recently, for more general cyclic models the Markov properties have been elucidated (Forré and Mooij, 2017).

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

When there are cyclic functional relationships between variables, one encounters various technical complications, which even arise in the linear setting. The structural equations of an acyclic SCM trivially have a unique solution. This unique solvability property ensures that the SCM gives rise to a unique, well-defined probability distribution on the variables. In the case of cycles, however, this property may be

violated, and consequently, the SCM may not have a solution at all, or may allow for multiple different probability distributions (Halpern, 1998). Even if one starts with a cyclic SCM that is uniquely solvable, performing an intervention on the SCM may lead to an intervened SCM that is not uniquely solvable. Hence, a cyclic SCM may not give rise to a unique, well-defined probability distribution corresponding to that intervention, and whether or not this happens may depend on the intervention. We provide sufficient conditions for the existence and uniqueness of these probability distributions after intervention. In general, it is not clear whether the solutions of the structural equations of an SCM are measurable if cycles are present. In addition, we provide sufficient and necessary conditions for the measurability of solution functions of cyclic SCMs.

SCMs provide a detailed modeling description of a system. Not all information may be necessary for a certain modeling task, which motivates to consider certain classes of SCMs to be equivalent. In this chapter, we formally introduce several of such equivalence relations. For example, we consider two SCMs observationally equivalent if they cannot be distinguished based on observations alone. Observationally equivalent SCMs can often still be distinguished by interventions. We consider two SCMs interventionally equivalent if they cannot be distinguished based on observations and interventions. While these concepts have been around in implicit form for acyclic SCMs, we formulate them in such a way that they also apply to cyclic SCMs that have either no solution at all or have multiple different induced probability distributions on the variables. Finally, we consider two SCMs counterfactually equivalent if they cannot be distinguished based on observations and interventions and in addition encode the same counterfactual distributions, which are the distributions induced by the so-called twin SCM via the twin network method (Balke and Pearl, 1994). These different equivalence relations formalize the different levels of abstraction in the so-called causal hierarchy (Shpitser and Pearl, 2008; Pearl and Mackenzie, 2018). In addition, we add another, strong version of equivalence, such that equivalent SCMs have the same solutions. This notion clarifies ambiguities when a function is constant in one of its arguments, for example.

Marginalization becomes useful if not all variables are observed: given a joint probability distribution on some variables, we obtain a marginal distribution on a subset of the variables by integrating out the remaining variables. Analogously, we can marginalize an acyclic SCM by substituting the solutions of the structural equations of a subset of the endogenous variables into the structural equations of the remaining endogenous variables. For acyclic SCMs, the induced observational and interventional distributions of the marginalized SCM coincide with the marginals of the distributions induced by the original SCM (see Verma, 1993; Spirtes et al., 1998; Evans, 2016; Evans, 2018; a.o.). In other words, for acyclic SCMs the operation of marginalization preserves the probabilistic and causal semantics (restricted to the remaining variables). We show that for cyclic SCMs a marginalization does not always exist without further assumptions. In (Forré and Mooij, 2017) it is shown that for modular SCMs, which can be seen as an SCM together with an additional

structure of a compatible system of solution functions, a marginalization can be defined that preserves the probabilistic and causal semantics. We prove that this additional structure is not necessary and use a local unique solvability condition instead. Under this condition, we show that an SCM and its marginalization are observationally, interventionally and counterfactually equivalent on the remaining endogenous variables. Analogously, we define a marginalization operation on the associated graph of an SCM, which generalizes the latent projection (Verma, 1993; Tian, 2002; Evans, 2016). In general, the marginalization of an SCM does not respect the latent projection of its associated graph, but we show that it does so under an additional local ancestral unique solvability condition.

In graphical models, Markov properties allow one to read off conditional independencies in a distribution directly from a graph. Various equivalent formulations of Markov properties exist for acyclic SCMs (Lauritzen, 1996), one prominent example being the d-separation criterion, also known as the directed global Markov property, which was originally derived for Bayesian networks (Pearl, 1985). Markov properties have been of key importance to derive various central results regarding causal reasoning and causal discovery. For cyclic SCMs, however, the usual Markov properties do not hold in general, as was already pointed out by Spirtes (1994). His solution in terms of collapsed graphs was recently generalized and reformulated for a general class of causal graphical models (Forré and Mooij, 2017) by adapting the notion of d-separation into what has been termed σ-separation. This resulted in a general directed global Markov property expressed in terms of σ-separation instead of d-separation. Here, we formulate these general Markov properties specifically within the framework of SCMs. Again, they only hold under certain unique solvability conditions.

In addition to its interpretation in terms of conditional independencies, the graph of an acyclic SCM also has a direct causal interpretation (Pearl, 2009). As was already observed in Neal (2000), the causal interpretation of SCMs with cycles can be counterintuitive, as the causal semantics under interventions no longer needs to be compatible with the structure imposed by the functional relations between the variables. We resolve this issue by showing that under certain ancestral unique solvability conditions the causal interpretation of SCMs is consistent with its graph.

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

directed global Markov property; their graphs have a direct and intuitive causal interpretation.

The scope of this chapter is limited to establishing the foundations for statistical causal modeling with cyclic SCMs (Figure 2.7 in Appendix 2.A.4 shows an overview of how SCMs relate to other causal graphical models). For a detailed discussion of causal reasoning, causal discovery and causal prediction with cyclic SCMs we refer the reader to other literature (e.g., Richardson, 1996a,b; Richardson and Spirtes, 1999; Eberhardt, Hoyer, and Scheines, 2010; Foygel, Draisma, and Drton, 2012; Hyttinen, Eberhardt, and Hoyer, 2012; Hyttinen et al., 2013). Several recent results (generalizations of the do-calculus, adjustment criteria and an identification algorithm) for modular SCMs (Forré and Mooij, 2018, 2019) directly apply to the subclass of simple SCMs, as well. Finally, many causal discovery algorithms that have been designed for the acyclic case also apply to simple SCMs with no or only minor changes (Mooij and Claassen, 2020; Mooij, Magliacane, and Claassen, 2020).

OVERVIEW Figure 2.1 gives an overview of the different objects that can be constructed from an SCM and the different mappings between them. For pairs of mappings between the objects with the names in bold, we prove commutativity results which are summarized in Table 2.1.

OUTLINE This chapter is structured as follows: In Section 2.2, we provide a formal definition of SCMs and a natural notion of equivalence between SCMs, define the (augmented) graph corresponding to an SCM, and describe perfect interventions and counterfactuals. In Section 2.3, we discuss the concept of (unique) solvability, its properties and how it relates to self-cycles. In Section 2.4, we define and relate various equivalence relations between SCMs. In Section 2.5, we define a marginalization operation that is applicable to cyclic SCMs under certain conditions. We discuss several properties of this marginalization operation and discuss the relation with a marginalization operation defined on directed mixed graphs. In Section 2.6, we discuss Markov properties of SCMs. In Section 2.7, we discuss the causal interpretation of the graphs of SCMs. Section 2.8 introduces and discusses the class of simple SCMs.

The appendices to this chapter introduces causal graphical models in Appendix 2.A. This section also contains details on Markov properties and modular SCMs. Appendix 2.B provides additional (unique) solvability properties, some results for linear SCMs are discussed in Appendix 2.C, other examples in Appendix 2.D and the proofs of all the theoretical results are in Appendix 2.E. Appendix 2.F contains some lemmas and measurable selection theorems that are used in several proofs.

2.2 STRUCTURAL CAUSAL MODELS

In this section, we provide the definition and properties of structural causal models (SCMs). Our definition of SCMs slightly deviates from existing definitions (Bollen, 1989; Spirtes, Glymour, and Scheines, 2000; Pearl, 2009), because we make the

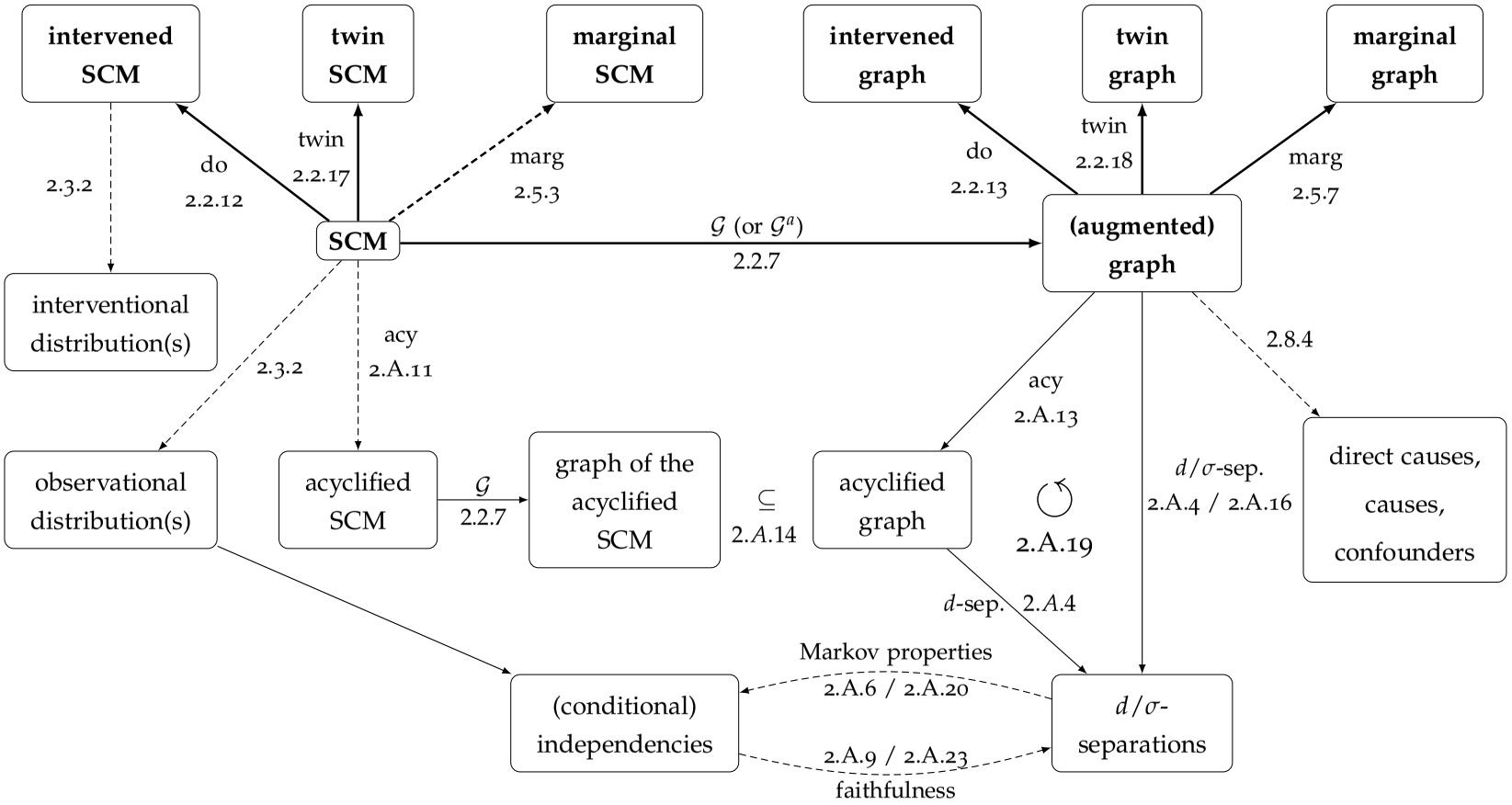


Figure 2.1: Overview of the objects constructed from an SCM and the mappings between them. The numbers correspond to the definition, proposition or theorem of the corresponding object, mapping or result. When an arrow is dashed, the relation only holds under nontrivial assumptions that can be found in the corresponding definition or theorem. The symbol "$\subseteq$" stands for the subgraph of a directed mixed graph (see Definition 2.A.1 in the Appendix 2.A) and the symbol "$\circlearrowleft$" denotes that the surrounding diagram commutes. Table 2.1 gives an overview of the commutativity results for each pair of mappings between the objects with the names in bold.

SCMs	do	twin	marg
$\mathcal{G}, \mathcal{G}^a$	2.2.14	2.2.19	(2.5.11)
do	2.2.15.(1)	2.2.21.(1)	2.5.5.(1)
twin	⋯	-	2.5.5.(2)
marg	⋯	⋯	2.5.4

Graphs	do	twin	marg
do	2.2.15.(1)	2.2.21.(2)	2.5.9.(1)
twin	⋯	-	2.5.9.(2)
marg	⋯	⋯	2.5.8

Table 2.1: Overview of the commutativity results of different pairs of mappings, defined on SCMs (left table) and on graphs (right table). All results apply under the assumptions stated in the corresponding proposition. The entries denoted by dots are omitted due to symmetry. We do not consider the commutativity of the twin operation with itself in this chapter. Proposition 2.5.11 (in parentheses) is not a commutativity result but a weaker relation. The graphical twin operator is only defined for directed graphs.

definition of the SCM independent of the random variables that solve it. This enables us to deal with the various technical complications that arise in the presence of cycles.

2.2.1 *Structural causal models and their solutions*

Definition 2.2.1 (Structural causal model)**.** *A* structural causal model (SCM) *is a tuple*[2]

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

where

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

In SCMs, the functional relationships between variables are expressed in terms of deterministic equations, where each equation expresses an endogenous variable (on the left-hand side) in terms of a causal mechanism depending on endogenous and

2 We often use boldface for variables that have multiple components, for example, vectors in a Cartesian product.

3 For the case $\mathcal{J} = \emptyset$, we have that $\boldsymbol{\mathcal{E}}$ is the singleton $\mathbf{1}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ is the degenerate probability measure $\mathbb{P}_{\mathbf{1}}$.

exogenous variables (on the right-hand side). This allows us to model interventions in an unambiguous way by changing the causal mechanisms that target specific endogenous variables (see Section 2.2.4).

Definition 2.2.2 (Structural equations)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call the set of equations*

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

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

Although it is common to assume the absence of cyclic functional relations (see Definition 2.2.9), we make no such assumption here. In particular, we allow for self-cycles, which we will discuss in more detail in Sections 2.2.2 and 2.3.3.

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

Definition 2.2.3 (Solution)**.** *A pair $(\boldsymbol{X}, \boldsymbol{E})$ of random variables $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}, \boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$, where Ω is a probability space, is a* solution *of the SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ if*

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

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

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

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

Example 2.2.4 (Cyclic SCMs)**.** *For brevity, we use throughout this chapter the notation $\boldsymbol{n} := \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$. Let $\mathcal{M} = \langle \boldsymbol{2}, \boldsymbol{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ be an SCM*[5] *with*

$$f_1(\boldsymbol{x}, e) = x_2\,,$$
$$f_2(\boldsymbol{x}, e) = x_1\,,$$

and $\mathbb{P}_{\mathbb{R}}$ an arbitrary probability measure on $\mathbb{R}$. Then (X, X) is a solution of $\mathcal{M}$ for any arbitrary random variable X with values in $\mathbb{R}$. Hence, any probability distribution on

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

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

$\{(x, x) : x \in \mathbb{R}\}$ *is an observational distribution associated to* $\mathcal{M}$*. Now consider instead the same SCM but with* $f_1(\boldsymbol{x}, e) = x_2 + 1$*. This SCM has no solutions at all, and hence induces no observational distribution.*

Due to the fact that the structural equations only need to be satisfied almost surely, there may exist many different SCMs representing the same set of solutions (see Example 2.D.4). It therefore seems natural not to differentiate between structural equations that have different solutions on at most a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set of exogenous variables. This leads to an equivalence relation between SCMs. To be able to state the equivalence relation concisely, we introduce the following notation: For subsets $\mathcal{U} \subseteq \mathcal{I}$ and $\mathcal{V} \subseteq \mathcal{J}$, we write $\boldsymbol{\mathcal{X}}_{\mathcal{U}} := \prod_{i \in \mathcal{U}} \mathcal{X}_i$ and $\boldsymbol{\mathcal{E}}_{\mathcal{V}} := \prod_{j \in \mathcal{V}} \mathcal{E}_j$. In particular, $\boldsymbol{\mathcal{X}}_{\emptyset}$ and $\boldsymbol{\mathcal{E}}_{\emptyset}$ are defined by the singleton $\mathbf{1}$. Moreover, for a subset $\mathcal{W} \subseteq \mathcal{I} \cup \mathcal{J}$, we use the convention that we write $\boldsymbol{\mathcal{X}}_{\mathcal{W}}$ and $\boldsymbol{\mathcal{E}}_{\mathcal{W}}$ instead of $\boldsymbol{\mathcal{X}}_{\mathcal{W} \cap \mathcal{I}}$ and $\boldsymbol{\mathcal{E}}_{\mathcal{W} \cap \mathcal{J}}$, respectively and we adopt a similar notation for the (random) variables in those spaces, that is, we write $\boldsymbol{x}_{\mathcal{W}}$ and $\boldsymbol{e}_{\mathcal{W}}$ instead of $\boldsymbol{x}_{\mathcal{W} \cap \mathcal{I}}$ and $\boldsymbol{e}_{\mathcal{W} \cap \mathcal{J}}$, respectively. This allows us to define the following natural equivalence relation for SCMs.[6,7]

Definition 2.2.5 (Equivalence)**.** *The two SCMs* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and* $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *are* equivalent*, denoted by* $\mathcal{M} \equiv \tilde{\mathcal{M}}$*, if for all* $i \in \mathcal{I}$*, for* $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$*-almost every* $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ *and for all* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*,*

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

Thus, two equivalent SCMs can only differ in terms of their causal mechanism. Importantly, equivalent SCMs have the same solutions and, as we will see in Sections 2.2.4 and 2.2.5, they have the same causal and counterfactual semantics (see Definitions 2.2.12 and 2.2.17, respectively). This equivalence relation on the set of all SCMs gives rise to the quotient set of equivalence classes of SCMs.

2.2.2 *The (augmented) graph*

We will now define two types of graphs that can be used for representing structural properties of the SCM. These graphical representations are related to Wright's path diagrams (Wright, 1921). The structural properties of the functional relations

6 An attempt at coarsening this notion of equivalence by replacing the quantifier "for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$" by "for almost every $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ under the observational distribution $\mathbb{P}^{\boldsymbol{X}}$" will not lead to a well-defined equivalence relation, since in general the observational distribution $\mathbb{P}^{\boldsymbol{X}}$ may be nonunique or even nonexistent. Refining it by replacing the quantifier "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" by "for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" would make it too fine for our purposes, since we assume the exogenous distribution to be fixed and we assume as usual that random variables that are almost surely identical are indistinguishable in practice. Note that the "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" and "for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$" quantifiers do not commute in general (see Example 2.D.5)

7 We may extend this definition to allow $\tilde{\mathcal{J}} \neq \mathcal{J}$ and for a larger class of SCMs such that the exogenous distribution does not factorize. Then, for any $\mathcal{M}$ that satisfies Definition 2.2.1, except for that it may have a non-factorizing exogenous distribution, there exists an equivalent SCM with a factorizing exogenous distribution (and a different $\mathcal{J}$); the latter can be obtained by partitioning the exogenous components into independent tuples. This motivates why we can restrict ourselves in Definition 2.2.1 to factorizing exogenous distributions only. For some more discussion on the representation of latent confounders, see also Example 2.D.6.

between variables modeled by an SCM are specified by the causal mechanism of the SCM and can be encoded in an (augmented) graph. For the graphical notation and standard terminology on directed (mixed) graphs that is used throughout this chapter, we refer the reader to Appendix 2.A.1.

We first define the parents of an endogenous variable.

Definition 2.2.6 (Parent). *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $k \in \mathcal{I} \cup \mathcal{J}$ a* parent of $i \in \mathcal{I}$ *if and only if there does not exist a measurable function*[8] *$\tilde{f}_i : \boldsymbol{\mathcal{X}}_{\setminus k} \times \boldsymbol{\mathcal{E}}_{\setminus k} \to \mathcal{X}_i$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,*

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

Exogenous variables have no parents by definition. These parental relations are preserved under the equivalence relation $\equiv$ on SCMs. They can be represented by a directed graph or a directed mixed graph.[9]

Definition 2.2.7 (Graph and augmented graph). *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We define:*

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

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

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

Example 2.2.8 (Graphs of an SCM). *Let $\mathcal{M} = \langle \mathbf{5}, \mathbf{3}, \mathbb{R}^5, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$ be an SCM with causal mechanism given by*

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

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

9 A *directed mixed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ consists of a set of nodes $\mathcal{V}$, a set of directed edges $\mathcal{E}$ and a set of bidirected edges $\mathcal{B}$ (see Definition 2.A.1 for a more precise definition).

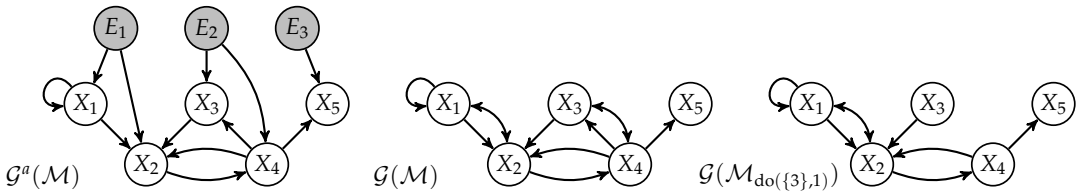


Figure 2.2: The augmented graph (left) and the graph (center) of the SCM $\mathcal{M}$ of Example 2.2.8 and the graph of the intervened SCM $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ of Example 2.2.16 (right).

where $\alpha \neq 0$ *and* $\mathbb{P}_{\mathbb{R}^3}$ *is a product of three probability measures* $\mathbb{P}_{\mathbb{R}}$ *over* $\mathbb{R}$ *that are non-degenerate. The augmented graph* $\mathcal{G}^a(\mathcal{M})$ *and the graph* $\mathcal{G}(\mathcal{M})$ *of* $\mathcal{M}$ *are depicted*[10] *in Figure 2.2 (left and center). Observe that if* α *had been equal to zero, then the endogenous variable* 1 *would not have any parents in* $\mathcal{G}^a(\mathcal{M})$*, that is, it would not have a self-cycle and directed edge from any exogenous variables in* $\mathcal{G}^a(\mathcal{M})$*, and it would not have a self-cycle and bidirected edge from any other variable in* $\mathcal{G}(\mathcal{M})$*. Moreover, if one of the probability measures* $\mathbb{P}_{\mathbb{R}}$ *over* $\mathbb{R}$ *were degenerate, then some of the directed edges from the exogenous variables to the endogenous variables in the augmented graph* $\mathcal{G}^a(\mathcal{M})$ *and bidirected edges in the graph* $\mathcal{G}(\mathcal{M})$ *would be missing.*

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

Definition 2.2.9. *We call an SCM* $\mathcal{M}$ acyclic *if* $\mathcal{G}^a(\mathcal{M})$ *is a directed acyclic graph (DAG). Otherwise, we call* $\mathcal{M}$ cyclic*.*

Equivalently, an SCM $\mathcal{M}$ is acyclic if $\mathcal{G}(\mathcal{M})$ is an acyclic directed mixed graph (ADMG) (Richardson, 2003). Acyclic SCMs are also known as semi-Markovian SCMs (Tian, 2002; Pearl, 2009). A commonly considered class of acyclic SCMs are the Markovian SCMs, which are acyclic SCMs for which each exogenous variable has at most one child. Several Markov properties were first shown for these models (Lauritzen et al., 1990; Tian, 2002; Pearl, 2009).

2.2.3 *Structurally minimal representations*

We have discussed an equivalence relation between SCMs in Section 2.2.1. In this subsection, we show that for each SCM there exists a representative of the equivalence class of that SCM for which each component of the causal mechanism does not depend on its nonparents (see also Peters, Janzing, and Schölkopf, 2017).

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

Definition 2.2.10 (Structurally minimal SCM). *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$* structurally minimal *if for all $i \in \mathcal{I}$ there exists a mapping $\tilde{f}_i : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(i)} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(i)} \to \mathcal{X}_i$ such that $f_i(\boldsymbol{x}, \boldsymbol{e}) = \tilde{f}_i(\boldsymbol{x}_{\mathrm{pa}(i)}, \boldsymbol{e}_{\mathrm{pa}(i)})$ for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.*

We already encountered a structurally minimal SCM $\mathcal{M}$ in Example 2.2.8. Taking instead $\alpha = 0$ in that example gives an SCM $\mathcal{M}$ that is not structurally minimal, since the endogenous variable 1 is then not a parent of itself, while $f_1(\boldsymbol{x}, \boldsymbol{e})$ depends on x_1. However, the equivalent SCM where we have replaced the causal mechanism of 1 by $f_1(\boldsymbol{x}, \boldsymbol{e}) = 0$ yields a structurally minimal SCM. In general, there always exists an equivalent structurally minimal SCM.

Proposition 2.2.11 (Existence of a structurally minimal SCM). *For an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, there exists an equivalent SCM $\tilde{\mathcal{M}} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ that is structurally minimal.*

For a causal mechanism $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ and a subset $\mathcal{U} \subseteq \mathcal{I}$, we write $\boldsymbol{f}_{\mathcal{U}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{U}}$ for the $\mathcal{U}$ components[11] of $\boldsymbol{f}$. A structurally minimal representation is compatible with the (augmented) graph, in the sense that for every $\mathcal{U} \subseteq \mathcal{I}$ there exists a unique measurable mapping $\tilde{\boldsymbol{f}}_{\mathcal{U}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{U})} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{U})} \to \boldsymbol{\mathcal{X}}_{\mathcal{U}}$ such that $\boldsymbol{f}_{\mathcal{U}}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_{\mathcal{U}}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{U})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{U})})$ for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. Moreover, for any $\mathcal{U} \subseteq \mathcal{I}$ there exists a unique measurable mapping $\tilde{\boldsymbol{f}}_{\mathrm{an}(\mathcal{U})} : \boldsymbol{\mathcal{X}}_{\mathrm{an}(\mathcal{U})} \times \boldsymbol{\mathcal{E}}_{\mathrm{an}(\mathcal{U})} \to \boldsymbol{\mathcal{X}}_{\mathrm{an}(\mathcal{U})}$ with $\boldsymbol{f}_{\mathrm{an}(\mathcal{U})}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{\boldsymbol{f}}_{\mathcal{U}}(\boldsymbol{x}_{\mathrm{an}(\mathcal{U})}, \boldsymbol{e}_{\mathrm{an}(\mathcal{U})})$ for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.

2.2.4 *Interventions*

To define the causal semantics of SCMs, we consider here an idealized class of interventions introduced by Pearl (2009) that we refer to as perfect interventions. Other types of interventions, like mechanism changes (Tian and Pearl, 2001), fat-hand interventions (Eaton and Murphy, 2007), activity interventions (Mooij and Heskes, 2013) and stochastic versions of all these are at least as relevant, but we do not consider them here.

Definition 2.2.12 (Perfect intervention on an SCM). *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value. The* perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *maps $\mathcal{M}$ to the SCM $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, where the* intervened causal mechanism $\tilde{\boldsymbol{f}}$ *is given by*

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

This operation $\mathrm{do}(I, \boldsymbol{\xi}_I)$ preserves the equivalence relation (see Definition 2.2.5) on the set of all SCMs, and hence this mapping induces a well-defined mapping on the set of equivalence classes of SCMs. Previous work has considered interventions

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

only on a specific subset of endogenous variables (Rubenstein et al., 2017; Beckers and Halpern, 2019; Blom, Bongers, and Mooij, 2019). Instead, we assume that we can intervene on any subset of endogenous variables in the model.

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

Definition 2.2.13 (Perfect intervention on a directed mixed graph)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $I \subseteq \mathcal{V}$ a subset. The perfect intervention $\mathrm{do}(I)$ maps $\mathcal{G}$ to the directed mixed graph $\mathrm{do}(I)(\mathcal{G}) := (\mathcal{V}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$, where $\tilde{\mathcal{E}} = \mathcal{E} \setminus \{v \to i : v \in \mathcal{V}, i \in I\}$ and $\tilde{\mathcal{B}} = \mathcal{B} \setminus \{v \leftrightarrow i : v \in \mathcal{V}, i \in \mathcal{I}\}$.*

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

Proposition 2.2.14. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value. Then $\big(\mathcal{G}^a \circ \mathrm{do}(I, \boldsymbol{\xi}_I)\big)(\mathcal{M}) = \big(\mathrm{do}(I) \circ \mathcal{G}^a\big)(\mathcal{M})$ and $\big(\mathcal{G} \circ \mathrm{do}(I, \boldsymbol{\xi}_I)\big)(\mathcal{M}) = \big(\mathrm{do}(I) \circ \mathcal{G}\big)(\mathcal{M})$.*

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

Proposition 2.2.15. *For an SCM and a directed mixed graph, we have the following properties:*

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

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

Example 2.2.16 (Intervened SCM and its graphs)**.** *Consider the SCM $\mathcal{M}$ of Example 2.2.8 which has a solution if and only if $\alpha \geq 0$. Applying the perfect intervention $\mathrm{do}(\{3\}, 1)$ to $\mathcal{M}$ gives the intervened model $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ with the intervened causal mechanism*

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

for which the graph $\mathcal{G}(\mathcal{M}_{\mathrm{do}(\{3\},1)})$ is depicted in Figure 2.2 (right). This is an example where a perfect intervention leads to an intervened SCM $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ that does not have a solution anymore. In addition, performing a perfect intervention $\mathrm{do}(\{4\}, 1)$ on $\mathcal{M}_{\mathrm{do}(\{3\},1)}$ yields again an SCM with a solution for $\alpha \geq 0$.

Recall that for each solution $\boldsymbol{X}$ of an SCM $\mathcal{M}$ we call the distribution $\mathbb{P}^{\boldsymbol{X}}$ the observational distribution of $\mathcal{M}$ associated to $\boldsymbol{X}$. For cyclic SCMs, the observational distribution is in general not unique.[12] For example, the SCM $\mathcal{M}$ of Example 2.2.8

12 In order to assure the existence of a unique observational distribution it is common to consider only SCMs for which the structural equations have a unique solution (see, e.g., Definition 7.1.1 in (Pearl, 2009)). Although these SCMs induce a unique observational distribution, they generally do not induce a unique distribution after a perfect intervention.

has two different observational distributions if $\alpha > 0$. Similarly, an intervened SCM may induce a distribution that is not unique. Whenever the intervened SCM $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ has a solution $\boldsymbol{X}$ we therefore call the distribution $\mathbb{P}^{\boldsymbol{X}}$ the *interventional distribution of $\mathcal{M}$ under the perfect intervention* $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *associated to* $\boldsymbol{X}$.[13]

2.2.5 *Counterfactuals*

The causal semantics of an SCM are described by the interventions on the SCM. Adding another layer of complexity, one can describe the counterfactual semantics of an SCM by the interventions on the so-called twin SCM, an idea introduced in Balke and Pearl (1994).

Definition 2.2.17 (Twin SCM)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. The* twin operation *maps $\mathcal{M}$ to the* twin structural causal model (twin SCM)

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

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

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

Definition 2.2.18 (Twin graph)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph and $\mathcal{I} \subseteq \mathcal{V}$ a subset such that $\mathcal{J} := \mathcal{V} \setminus \mathcal{I}$ is* exogenous*, that is, $\mathrm{pa}_{\mathcal{G}}(\mathcal{J}) = \emptyset$. The* $\mathrm{twin}(\mathcal{I})$ operation *maps $\mathcal{G}$ to the* twin graph w.r.t. $\mathcal{I}$ *defined by* $\mathrm{twin}(\mathcal{I})(\mathcal{G}) := (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$*, where:*

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

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

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

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

Proposition 2.2.19. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. Then $(\mathcal{G}^a \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin}(\mathcal{I}) \circ \mathcal{G}^a)(\mathcal{M})$.*

Proposition 2.2.20. *For SCMs and directed graphs, we have that acyclicity is preserved under the twin operation.*

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

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

Proposition 2.2.21. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM and $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a directed graph. Then we have that perfect intervention commutes with the twin operation on both:*

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

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

Whenever the intervened twin SCM $(\mathcal{M}^{\mathrm{twin}})_{\mathrm{do}(\tilde{I}, \boldsymbol{\xi}_{\tilde{I}})}$, where $\tilde{I} \subseteq \mathcal{I} \cup \mathcal{I}'$ and $\boldsymbol{\xi}_{\tilde{I}} \in \boldsymbol{\mathcal{X}}_{\tilde{I}}$, has a solution $(\boldsymbol{X}, \boldsymbol{X}')$, we call the distribution $\mathbb{P}^{(\boldsymbol{X}, \boldsymbol{X}')}$ the *counterfactual distribution of $\mathcal{M}$ under the perfect intervention* $\mathrm{do}(\tilde{I}, \boldsymbol{\xi}_{\tilde{I}})$ *associated to* $(\boldsymbol{X}, \boldsymbol{X}')$. In Example 2.D.3, we provide an example of how counterfactuals can be sensibly formulated for a well-known market equilibrium model described in terms of a cyclic SCM.

The interpretation of counterfactual statements has received a lot of attention in the literature (Lewis, 1979; Balke and Pearl, 1994; Roese, 1997; Byrne, 2007; Pearl, 2009). For acyclic graphs, an alternative graphical approach to counterfactuals is the framework of Single World Intervention Graphs (SWIGs) (Richardson and Robins, 2013). One topic of discussion is that there exist SCMs that induce the same observational and interventional distributions, but differ in their counterfactual statements (Dawid, 2002) (see also Example 2.D.7). This raises the question how one can estimate such SCMs from data.

2.3 SOLVABILITY

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

2.3.1 *Definition of solvability*

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

Definition 2.3.1 (Solvability)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$* solvable w.r.t. *$\mathcal{O} \subseteq \mathcal{I}$ if there exists a measurable mapping $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,*

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

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

By definition, solvability w.r.t. a subset respects the equivalence relation $\equiv$ on SCMs. The measurable solution functions w.r.t. a certain subset do not always exist, and if they exist, they are not always uniquely defined. For example, for the SCM $\mathcal{M}$ in Example 2.2.8, the measurable solution functions w.r.t. $\{1\}$ are given by $g_1^{\pm}(e_1) = \pm\sqrt{\alpha e_1^2}$ if and only if $\alpha \geq 0$.

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

Theorem 2.3.2 (Sufficient and necessary conditions for solvability)**.** *For an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, the following are equivalent:*

1. *$\mathcal{M}$ has a solution (see Definition 2.2.3);*
2. *for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ the structural equations $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{e})$ have a solution $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$;*
3. *$\mathcal{M}$ is solvable (see Definition 2.3.1).*

While in the acyclic case, the above theorem is almost trivial, in the cyclic case the measure-theoretic aspects are not that obvious. In particular, to prove the existence of a *measurable* solution function $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ in case the structural equations have a solution for almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, we make use of a strong measurable selection theorem (see Theorem 2.F.8 or (Kechris, 1995)). This theorem implies that if there exists a solution $\boldsymbol{X} : \Omega \to \boldsymbol{\mathcal{X}}$, then there necessarily exists a random variable $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$ and a mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{g}(\boldsymbol{E}_{\mathrm{pa}(\mathcal{I})})$ is a solution. However, it does not imply that there necessarily exists a random variable $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$ and a mapping $\boldsymbol{g} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{E}_{\mathrm{pa}(\mathcal{I})})$ holds a.s., for example, if $\boldsymbol{X}$ is a nontrivial mixture of such solutions (see Example 2.D.8).

Solvability w.r.t. a strict subset of $\mathcal{I}$ is in general neither sufficient nor necessary for the existence of a (global) solution of the SCM. Consider, for example, the SCM $\mathcal{M}$ in Example 2.2.8 with $\alpha < 0$. Even though this SCM is solvable w.r.t. $\{2, 3, 4\}$, it is not (globally) solvable, and hence does not have any solution. In Proposition 2.B.1, we provide a sufficient condition for solvability w.r.t. a strict subset of $\mathcal{I}$ that is similar to condition (2) in Theorem 2.3.2 in the sense that it is formulated in terms of the solutions of (a subset of) the structural equations without requiring measurability of the solutions. For the class of linear SCMs, we provide in Proposition 2.C.2 a sufficient and necessary condition for solvability w.r.t. a subset of $\mathcal{I}$.

Figure 2.3: Left: The graphs of the observationally equivalent SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ of Examples 2.3.5 and 2.4.2, respectively. Right: The graphs of the interventionally equivalent SCMs $\bar{\mathcal{M}}$ and $\hat{\mathcal{M}}$ of Example 2.4.4.

2.3.2 *Unique solvability*

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

Definition 2.3.3 (Unique solvability)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$* uniquely solvable w.r.t. *$\mathcal{O} \subseteq \mathcal{I}$ if there exists a measurable mapping $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,*

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

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

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

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

Proposition 2.3.4. *An acyclic SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is uniquely solvable w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$.*

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

Example 2.3.5 (Cyclic SCM, uniquely solvable w.r.t. each subset)**.** *Consider the SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with $\mathcal{I} = \mathcal{J} = \mathbf{4}$, $\mathcal{X}_i = \mathcal{E}_i = (-1, 1)$ for $i = 1, 2$, and $\mathcal{X}_i = \mathcal{E}_i = \mathbb{R}$ for $i = 3, 4$, the causal mechanism given by*

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

and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ the standard-normal distribution on $\mathbb{R}^4$ restricted[14] to $\boldsymbol{\mathcal{E}}$. This SCM $\mathcal{M}$ is uniquely solvable w.r.t. every subset and its (augmented) graph includes a cycle (see Figure 2.3).

14 The *restriction* of a probability measure $\mathbb{P}$, of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, on a measurable subset $\mathcal{O} \in \mathcal{F}$ is the *restricted* probability measure $\mathbb{P}_{\mathcal{O}} := \frac{\mathbb{P}|_{\mathcal{F}_{\mathcal{O}}}}{\mathbb{P}(\mathcal{O})}$, of the probability space $(\mathcal{O}, \mathcal{F}_{\mathcal{O}}, \mathbb{P}_{\mathcal{O}})$, where $\mathcal{F}_{\mathcal{O}} := \{\mathcal{V} \cap \mathcal{O} : \mathcal{V} \in \mathcal{F}\}$.

Theorem 2.3.2 provides sufficient and necessary conditions for (global) solvability. The next theorem states that under the additional uniqueness requirement there exists a sufficient and necessary condition for unique solvability w.r.t. any subset (for solvability w.r.t. a subset we only have the sufficient condition provided in Proposition 2.B.1), and moreover, that all solutions of a uniquely solvable SCM induce the same observational distribution.

Theorem 2.3.6 (Sufficient and necessary conditions for unique solvability)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM and $\mathcal{O} \subseteq \mathcal{I}$ a subset. The following are equivalent:*

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

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

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

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

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

It is well known that under acyclicity the observational distribution is unique. Theorem 2.3.6 generalizes this result to settings with cycles. For linear SCMs, the unique solvability condition w.r.t. a subset is equivalent to a matrix invertibility condition (see Proposition 2.C.3).

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

2.3.3 *Self-cycles*

One can think of a structural equation of a single endogenous variable $i \in \mathcal{I}$ as describing a small subsystem that interacts with the rest of the system. If the output x_i of this subsystem is uniquely determined by the input $(\boldsymbol{x}_{\setminus i}, \boldsymbol{e})$ from the rest of the system (up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set), then i is not a parent of itself (see Definition 2.2.6).

Proposition 2.3.7 (Self-cycles)**.** *The SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is uniquely solvable w.r.t. $\{i\}$ for $i \in \mathcal{I}$ if and only if $\mathcal{G}^a(\mathcal{M})$ (or $\mathcal{G}(\mathcal{M})$) has no self-cycle $i \to i$ at $i \in \mathcal{I}$.*

A self-cycle at an endogenous variable denotes that that variable is not uniquely determined by its parents, up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set. This implies that an SCM with a self-cycle at an endogenous variable in its graph can be either solvable, or not solvable, w.r.t. that variable. For the SCM $\mathcal{M}$ of Example 2.2.8, we have indeed that it is solvable w.r.t. $\{1\}$ for $\alpha > 0$, while for $\alpha < 0$ it is not. For linear SCMs with structural equations $X_i = \sum_{j \in \mathcal{I}} B_{ij} X_j + \sum_{k \in \mathcal{J}} \Gamma_{ik} E_k$, the endogenous variable $i \in \mathcal{I}$ has a self-cycle if and only if $B_{ii} = 1$ (see also Appendix 2.C).

2.3.4 *Interventions*

The property of (unique) solvability is in general not preserved under perfect intervention. For example, a (uniquely) solvable SCM can lead to a nonuniquely solvable SCM after intervention, which either has no solution or has solutions with multiple induced distributions (see, e.g., Examples 2.2.16 and 2.D.9). A sufficient condition for the intervened SCM to be (uniquely) solvable is that the original SCM has to be (uniquely) solvable w.r.t. the subset of nonintervened endogenous variables.

Proposition 2.3.8. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is (uniquely) solvable w.r.t. $\mathcal{O} \subseteq \mathcal{I}$. Then, for any set I such that $\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O} \subseteq I \subseteq \mathcal{I} \setminus \mathcal{O}$ and value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ the intervened SCM $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is (uniquely) solvable w.r.t. $\mathcal{O} \cup I$.*

Proposition 2.3.4 shows that acyclic SCMs are uniquely solvable w.r.t. every subset and hence are uniquely solvable after every perfect intervention. This also directly follows from the fact that acyclicity is preserved under perfect intervention (see Proposition 2.2.15). Moreover, since acyclicity is preserved under the twin operation (see Proposition 2.2.20), an acyclic SCM induces unique observational, interventional and counterfactual distributions.

2.3.5 *Ancestral (unique) solvability*

We saw that, in general, solvability w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply solvability w.r.t. a strict subset of $\mathcal{O}$. Here we show that it does imply solvability w.r.t. the ancestral subsets in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$, that is, in the induced subgraph of the graph $\mathcal{G}(\mathcal{M})$ on $\mathcal{O}$. A subset $\mathcal{A} \subseteq \mathcal{O}$ is called an *ancestral subset* in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$ if $\mathcal{A} = \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$, where $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$ are the ancestors of $\mathcal{A}$ according to the induced subgraph[15] $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$.

Definition 2.3.9 (Ancestral (unique) solvability)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$* ancestrally (uniquely) solvable w.r.t. *$\mathcal{O} \subseteq \mathcal{I}$ if $\mathcal{M}$ is (uniquely) solvable w.r.t. every ancestral subset in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$. We call $\mathcal{M}$* ancestrally (uniquely) solvable *if it is ancestrally (uniquely) solvable w.r.t. $\mathcal{I}$.*

Proposition 2.3.10 (Solvability is equivalent to ancestral solvability)**.** *The SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ is solvable w.r.t. the subset $\mathcal{O} \subseteq \mathcal{I}$ if and only if $\mathcal{M}$ is ancestrally solvable w.r.t. $\mathcal{O}$.*

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

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

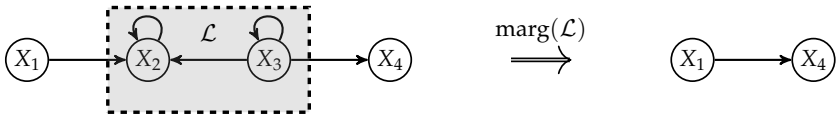


Figure 2.4: The graphs of the SCM $\mathcal{M}$ (left) of Example 2.3.11 and the marginal SCM $\mathcal{M}_{\text{marg}(\{2,3\})}$ (right) of Example 2.5.10.

Example 2.3.11 (Unique solvability w.r.t. $\mathcal{O}$ does not imply ancestral unique solvability w.r.t. $\mathcal{O}$)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{4}, \mathbf{1}, \mathbb{R}^4, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ *with causal mechanism given by*

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

and $\mathbb{P}_{\mathbb{R}}$ *the standard-normal measure on* $\mathbb{R}$*. This SCM is uniquely solvable w.r.t. the set* $\{2,3\}$*, and thus solvable w.r.t. this set. Although it is solvable w.r.t., the ancestral subset* $\{3\}$ *in* $\mathcal{G}(\mathcal{M})_{\{2,3\}}$*, depicted in Figure 2.4 (left), it is not uniquely solvable w.r.t. this subset, because the structural equation* $x_3 = x_3$ *holds for any* $x_3 \in \mathbb{R}$*. Hence, it is not ancestrally uniquely solvable w.r.t.* $\{2,3\}$*.*

However, for the class of linear SCMs we have that unique solvability w.r.t. $\mathcal{O}$ always implies ancestral unique solvability w.r.t. $\mathcal{O}$ (see Proposition 2.C.4).

Although in general unique solvability is not preserved under unions, in Proposition 2.B.4 we show that if an SCM is uniquely solvable w.r.t. two ancestral subsets and w.r.t. their intersection, then it is uniquely solvable w.r.t. their union. In general, the property of ancestral unique solvability is not preserved under perfect intervention, as can be seen in Example 2.D.9. The notion of ancestral unique solvability will appear in various results in Sections 2.5 and 2.6.

2.4 EQUIVALENCES

In Section 2.2, we already encountered an equivalence relation on the class of SCMs (see Definition 2.2.5). The (augmented) graph of an SCM, its solutions and its induced observational, interventional and counterfactual distributions are preserved under this equivalence relation. In this section, we give several coarser equivalence relations on the class of SCMs: observational, interventional and counterfactual equivalence.

2.4.1 *Observational equivalence*

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

Definition 2.4.1 (Observational equivalence)**.** *Two SCMs* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and* $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ *are* observationally equivalent w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$*, denoted by* $\mathcal{M} \equiv_{obs(\mathcal{O})} \tilde{\mathcal{M}}$*, if* $\boldsymbol{\mathcal{X}}_{\mathcal{O}} = \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{O}}$ *and for all solutions* $\boldsymbol{X}$ *of* $\mathcal{M}$ *there exists a solution* $\tilde{\boldsymbol{X}}$

of $\tilde{\mathcal{M}}$ such that $\mathbb{P}^{\boldsymbol{X}_{\mathcal{O}}} = \mathbb{P}^{\tilde{\boldsymbol{X}}_{\mathcal{O}}}$ and for all solutions $\tilde{\boldsymbol{X}}$ of $\tilde{\mathcal{M}}$ there exists a solution $\boldsymbol{X}$ of $\mathcal{M}$ such that $\mathbb{P}^{\boldsymbol{X}_{\mathcal{O}}} = \mathbb{P}^{\tilde{\boldsymbol{X}}_{\mathcal{O}}}$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are called observationally equivalent *if they are observationally equivalent w.r.t. $\mathcal{I} = \tilde{\mathcal{I}}$.*

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

Example 2.4.2 (Observational equivalence does not imply equivalence)**.** *Consider the SCM $\tilde{\mathcal{M}}$ that is the same as $\mathcal{M}$ of Example 2.3.5 but with the causal mechanism $\tilde{\boldsymbol{f}}$ given by*

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

This SCM $\tilde{\mathcal{M}}$ is observationally equivalent to the SCM $\mathcal{M}$. Because both SCMs have a different (augmented) graph they are not equivalent to each other (see Figure 2.3).

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

2.4.2 *Interventional equivalence*

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

Definition 2.4.3 (Interventional equivalence)**.** *Two SCMs $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ and $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ are* interventionally equivalent w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$*, denoted by $\mathcal{M} \equiv_{int(\mathcal{O})} \tilde{\mathcal{M}}$, if $\boldsymbol{\mathcal{X}}_{\mathcal{O}} = \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{O}}$ and for every $I \subseteq \mathcal{O}$ and every value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ their intervened models $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ and $\tilde{\mathcal{M}}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ are observationally equivalent with respect to $\mathcal{O}$. $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are called* interventionally equivalent *if they are interventionally equivalent w.r.t. $\mathcal{I} = \tilde{\mathcal{I}}$.*

Equivalent SCMs have the same solutions under every perfect intervention, and hence they are interventionally equivalent w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$. SCMs that are interventionally equivalent w.r.t. a subset $\mathcal{O} \subseteq \mathcal{I}$ are interventionally equivalent w.r.t. every strict subset $\mathcal{W} \subsetneq \mathcal{O}$. But in general, they are not interventionally equivalent w.r.t. a strict superset $\mathcal{O} \subsetneq \mathcal{V} \subseteq \mathcal{I}$, as can be seen in Example 2.4.2, where the SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are interventionally equivalent w.r.t. $\{1, 2\}$ but are not interventionally equivalent. Interventional equivalence w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ implies observational equivalence w.r.t. $\mathcal{O}$, since the empty perfect intervention ($I = \emptyset$) is a special case of a perfect intervention. However, observational equivalence w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply interventional equivalence w.r.t. $\mathcal{O}$ in general, as can be seen in Example 2.4.2, where the SCMs $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are observationally equivalent but not interventionally equivalent.

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

Example 2.4.4 (Interventionally equivalent SCMs with different graphs)**.** *Consider the SCM* $\bar{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \{-1,1\}^2, \{-1,1\}^2, \bar{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and the SCM* $\hat{\mathcal{M}}$ *that is the same as* $\bar{\mathcal{M}}$ *except for its causal mechanism* $\hat{\boldsymbol{f}}$*, where the causal mechanisms are given by*

$$\begin{aligned} \bar{f}_1(\boldsymbol{x}, \boldsymbol{e}) &= e_1\,, & \hat{f}_1(\boldsymbol{x}, \boldsymbol{e}) &= e_1\,, \\ \bar{f}_2(\boldsymbol{x}, \boldsymbol{e}) &= x_1 e_2\,, & \hat{f}_2(\boldsymbol{x}, \boldsymbol{e}) &= e_2\,, \end{aligned}$$

and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^E$ *with* $E_1, E_2 \sim \mathcal{U}(\{-1,1\})$ *uniformly distributed and* $E_1 \perp\!\!\!\perp E_2$*. Then* $\bar{\mathcal{M}}$ *and* $\hat{\mathcal{M}}$ *are interventionally equivalent although* $\mathcal{G}(\bar{\mathcal{M}})$ *is not equal to* $\mathcal{G}(\hat{\mathcal{M}})$ *(see Figure 2.3).*

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

2.4.3 *Counterfactual equivalence*

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

Definition 2.4.5 (Counterfactual equivalence)**.** *Two SCMs* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *and* $\tilde{\mathcal{M}} = \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$ *are* counterfactually equivalent with respect to $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$*, denoted by* $\mathcal{M} \equiv_{cf(\mathcal{O})} \tilde{\mathcal{M}}$*, if* $\mathcal{M}^{\text{twin}}$ *and* $\tilde{\mathcal{M}}^{\text{twin}}$ *are interventionally equivalent with respect to* $\mathcal{O} \cup \mathcal{O}'$*, where* $\mathcal{O}'$ *corresponds to the copy of* $\mathcal{O}$ *in* $\mathcal{I}' \cap \tilde{\mathcal{I}}'$*.* $\mathcal{M}$ *and* $\tilde{\mathcal{M}}$ *are called* counterfactually equivalent *if they are counterfactually equivalent with respect to* $\mathcal{I} = \tilde{\mathcal{I}}$*.*

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

Proposition 2.4.6. *For SCMs, we have that equivalence implies counterfactual equivalence w.r.t.* $\mathcal{O}$*, which in turn implies interventional equivalence w.r.t.* $\mathcal{O}$*, for any* $\mathcal{O} \subseteq \mathcal{I}$*.*

Interventionally equivalent SCMs that have the same causal mechanism (that differ only in their exogenous distribution) may not be counterfactually equivalent (see, e.g., Example 2.D.7). Although the notion of counterfactual equivalence is finer than the notion of observational and interventional equivalence, the (augmented) graphs for counterfactually equivalent SCMs are in general not equal to each other (see Example 2.D.10).

2.4.4 *Relations between equivalences*

The definitions of observational, interventional and counterfactual equivalence provide equivalence relations on the set of all SCMs. For two SCMs to be observationally, interventionally or counterfactually equivalent w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$, the domains of their endogenous variables $\mathcal{O}$ have to be equal, that is, $\boldsymbol{\mathcal{X}}_{\mathcal{O}} = \tilde{\boldsymbol{\mathcal{X}}}_{\mathcal{O}}$. Apart from that, the index sets of the endogenous and the exogenous variables, the spaces of the other endogenous and exogenous variables, the causal mechanism and the exogenous probability measure may all differ. The observational, interventional and counterfactual equivalence classes w.r.t. $\mathcal{O} \subseteq \mathcal{I} \cap \tilde{\mathcal{I}}$ are related in the following way (see Proposition 2.4.6):

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

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

2.5 MARGINALIZATIONS

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

The problem of marginalization of directed graphical models has been addressed for acyclic graph structures, for example, ADMGs and mDAGs (see Verma, 1993; Richardson and Spirtes, 2002; Richardson, 2003; Evans, 2016; Evans, 2018; a.o.), and more recently in (Forré and Mooij, 2017) for certain graph structures ("HEDGes") that may include cycles. Although in the acyclic setting it has been shown that the marginalization for some of these graph structures preserves the probabilistic and causal semantics, in the cyclic setting this has only been shown for modular SCMs (Forré and Mooij, 2017). We show that without the additional structure of a compatible system of solution functions (see Appendix 2.A.3) one can still define a marginalization for SCMs under certain local unique solvability conditions. Intuitively, the idea is that if the state of a subsystem of endogenous variables is uniquely determined by the parents outside of this subsystem, then one can ignore the internals of this subsystem by treating it as a "black box" that can be described by certain measurable solution functions (see Figure 2.4). One can marginalize over this subsystem by substituting these measurable solution functions into the rest of the model, thereby removing the functional dependencies on the variables

of the subsystem from the rest of the system, while preserving the probabilistic, causal and the counterfactual semantics of the rest of the system. We show that in general this marginalization operation defined on SCMs does not respect the latent projection on its associated (augmented) graph, where the latent projection is a similar marginalization operation defined on directed mixed graphs (Verma, 1993; Tian, 2002; Evans, 2016). We show that under certain stronger local ancestral unique solvability conditions the marginalization does respect the latent projection.

2.5.1 *Marginalization of a structural causal model*

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

Example 2.5.1 (No SCM on the margin preserves the causal semantics)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *with causal mechanism*

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

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

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

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

Example 2.5.2 (SCM on the margin that preserves the causal semantics)**.** *Consider the SCM* $\mathcal{M}$ *of Example 2.3.11 that is uniquely solvable w.r.t. the subset* $\mathcal{L} = \{2,3\}$ *(depicted by the gray box in Figure 2.4). Substituting the measurable solution functions* $\boldsymbol{g}_{\mathcal{L}}$ *into the causal mechanism components* f_1 *and* f_4 *for the remaining endogenous variables* $\{1,4\}$ *gives a "marginal" causal mechanism* $\tilde{f}_1(\boldsymbol{x}, e) := e$ *and* $\tilde{f}_4(\boldsymbol{x}, e) := x_1$*. This defines an SCM* $\tilde{\mathcal{M}}$ *on the margin* $\mathcal{I} \setminus \mathcal{L} = \{1,4\}$ *that is interventionally equivalent w.r.t.* $\mathcal{I} \setminus \mathcal{L}$ *to* $\mathcal{M}$*.*

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

the model by substitution. This leads to a marginal SCM that is observationally, interventionally and counterfactually equivalent to the original SCM w.r.t. the margin, as we prove in Theorem 2.5.6.

Definition 2.5.3 (Marginalization of an SCM)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$ and let $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$. For $\boldsymbol{g}_{\mathcal{L}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L})\setminus\mathcal{L}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L})} \to \boldsymbol{\mathcal{L}}$, any measurable solution function of $\mathcal{M}$ w.r.t. $\mathcal{L}$, we call the SCM $\mathcal{M}_{\mathrm{marg}(\mathcal{L})} := \langle \mathcal{O}, \mathcal{J}, \boldsymbol{\mathcal{X}}_{\mathcal{O}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with the* marginal causal mechanism *$\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}}_{\mathcal{O}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ given by*

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

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

The marginalization of $\mathcal{M}$ w.r.t. $\mathcal{L}$ is defined up to the equivalence $\equiv$ on SCMs, since the measurable solution functions $\boldsymbol{g}_{\mathcal{L}}$ are uniquely defined up to $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null sets. With this definition at hand, we can always construct a marginal SCM over a subset of the endogenous variables of an acyclic SCM by mere substitution (see also Proposition 2.3.4). Moreover, this definition extends that notion to SCMs that are uniquely solvable w.r.t. a certain subset. For linear SCMs this condition translates into a matrix invertibility condition, and since substitution preserves linearity, marginalization yields a linear marginal SCM (see Proposition 2.C.5).

In general, marginalization is not always defined for all subsets. For instance, the SCM of Example 2.3.11 cannot be marginalized over the variable 3 (due to the self-cycle at 3), but can be marginalized over the variables 2 and 3 together. It follows from Proposition 2.3.7 that we can only marginalize over a single variable if that variable has no self-cycle. Note that we may introduce new self-cycles if we marginalize over a subset of variables, as can be seen, for example, from the SCM $\mathcal{M}$ in Example 2.2.8. This SCM has only one self-cycle; however, marginalizing w.r.t. $\{2\}$ gives a marginal SCM with another self-cycle at variable 4.

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

Proposition 2.5.4. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L}_1 \subseteq \mathcal{I}$ and let $\mathcal{L}_2 \subseteq \mathcal{I}$ be a subset disjoint from $\mathcal{L}_1$. Then $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$ if and only if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}_1 \cup \mathcal{L}_2$, Moreover,*

$$\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1)(\mathcal{M}) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{M}) \,.$$

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

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

Proposition 2.5.5. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$. Then the marginalization* $\mathrm{marg}(\mathcal{L})$ *commutes with both:*

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

2. *the twin operation* twin, *that is,*
$$(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin})(\mathcal{M}) = (\mathrm{twin} \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M}),$$

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

With Proposition 2.5.5 at hand, we can prove the main result of this subsection.

Theorem 2.5.6 (Marginalization of an SCM preserves the observational, causal and counterfactual semantics)**.** *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. a subset $\mathcal{L} \subseteq \mathcal{I}$. Then $\mathcal{M}$ and* $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ *are observationally, interventionally and counterfactually equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.*

This shows that our definition of marginalization (Definition 2.5.3) preserves the probabilistic, causal and counterfactual semantics, under a certain local unique solvability condition. Moreover, this allows us to marginalize SCMs w.r.t. a certain subset that do not satisfy the additional assumptions imposed by modular SCMs, for example, the SCM $\mathcal{M}$ of Example 2.3.11 does not have any additional structure of a compatible system of solution functions, but $\mathcal{M}$ can be marginalized w.r.t. the subset $\{2, 3\}$ (see also Appendix 2.A.3).

In general, interventional equivalence does not imply counterfactual equivalence (see, e.g., Example 2.D.7). However, for our definition of marginalization we arrive at a marginal SCM that is not only interventionally equivalent, but also counterfactually equivalent w.r.t. the margin.

For an SCM $\mathcal{M}$, unique solvability w.r.t. a certain subset $\mathcal{L} \subseteq \mathcal{I}$ is a sufficient, but not a necessary condition for the existence of an SCM $\tilde{\mathcal{M}}$ on the margin $\mathcal{I} \setminus \mathcal{L}$ such that $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are counterfactually equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$ (see, e.g., Example 2.D.11). Hence, in certain cases it may be possible to relax the uniqueness condition.

2.5.2 *Marginalization of a graph*

We now turn to a marginalization operation for directed mixed graphs, which we call the latent projection. This name is inspired from a similar construction on directed mixed graphs in (Verma, 1993). In (Verma, 1993), the authors concentrate on a mapping between directed mixed graphs and show that it preserves conditional independence properties (see also Tian, 2002). In this subsection, we provide a sufficient condition for the marginalization of an SCM to respect the latent

projection, that is, that the augmented graph of the marginal SCM is a subgraph of the latent projection of the augmented graph of the original SCM.

Definition 2.5.7 (Marginalization/latent projection of a directed mixed graph)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathcal{L} \subseteq \mathcal{V}$ a subset. The* marginalization of $\mathcal{G}$ w.r.t. $\mathcal{L}$ *or the* latent projection of $\mathcal{G}$ onto $\mathcal{V} \setminus \mathcal{L}$ *maps $\mathcal{G}$ to the* marginal graph $\mathrm{marg}(\mathcal{L})(\mathcal{G}) := (\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$*, where:*

1. $\tilde{\mathcal{V}} = \mathcal{V} \setminus \mathcal{L}$,
2. *for $i, j \in \tilde{\mathcal{V}}$: $i \to j \in \tilde{\mathcal{E}}$ if and only if there exists a directed path*

$$i \to \ell_1 \to \cdots \to \ell_n \to j$$

   in $\mathcal{G}$ with $n \geq 0$ and $\ell_1, \ldots, \ell_n \in \mathcal{L}$,
3. *for $i \neq j \in \tilde{\mathcal{V}}$: $i \leftrightarrow j \in \tilde{\mathcal{B}}$ if and only if*
   a) *there exist $n, m \geq 0$, $\ell_1, \ldots, \ell_n \in \mathcal{L}$, $\tilde{\ell}_1, \ldots, \tilde{\ell}_m \in \mathcal{L}$ such that*

$$i \leftarrow l_1 \leftarrow l_2 \leftarrow \cdots \leftarrow \ell_n \leftrightarrow \tilde{\ell}_m \to \tilde{\ell}_{m-1} \to \cdots \to \tilde{\ell}_1 \to j$$

      in $\mathcal{G}$, or
   b) *there exist $n, m \geq 1$, $\ell_1, \ldots, \ell_n \in \mathcal{L}$, $\tilde{\ell}_1, \ldots, \tilde{\ell}_m \in \mathcal{L}$ such that*

$$i \leftarrow l_1 \leftarrow l_2 \leftarrow \cdots \leftarrow \ell_n$$

      and

$$\tilde{\ell}_m \to \tilde{\ell}_{m-1} \to \cdots \to \tilde{\ell}_1 \to j$$

      in $\mathcal{G}$ and $\ell_n = \tilde{\ell}_m$.

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

Proposition 2.5.8. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{V}$ two disjoint subsets. Then*

$$(\mathrm{marg}(\mathcal{L}_1) \circ \mathrm{marg}(\mathcal{L}_2))(\mathcal{G}) = (\mathrm{marg}(\mathcal{L}_2) \circ \mathrm{marg}(\mathcal{L}_1))(\mathcal{G}) = \mathrm{marg}(\mathcal{L}_1 \cup \mathcal{L}_2)(\mathcal{G})\,.$$

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

Proposition 2.5.9. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathcal{L}, \mathcal{I}, I \subseteq \mathcal{V}$ subsets. Then the marginalization* $\mathrm{marg}(\mathcal{L})$ *commutes with both:*

1. *perfect intervention* $\mathrm{do}(I)$ *if I is disjoint from $\mathcal{L}$, that is,*

$$(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I))(\mathcal{G}) = (\mathrm{do}(I) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{G})\,,$$

and

2. *the twin operation* $\text{twin}(\mathcal{I})$ *if* $\mathcal{B} = \emptyset$, $\mathcal{J} := \mathcal{V} \setminus \mathcal{I}$ *is exogenous (i.e.,* $\text{pa}_{\mathcal{G}}(\mathcal{J}) = \emptyset$*) and* $\mathcal{L} \subseteq \mathcal{I}$*, that is,*

$$(\text{marg}(\mathcal{L} \cup \mathcal{L}') \circ \text{twin}(\mathcal{I}))(\mathcal{G}) = (\text{twin}(\mathcal{I} \setminus \mathcal{L}) \circ \text{marg}(\mathcal{L}))(\mathcal{G}),$$

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

An example of an SCM for which a marginalization respects the latent projection is the SCM $\mathcal{M}$ of Example 2.2.8. Marginalizing $\mathcal{M}$ w.r.t. $\mathcal{L} = \{2\}$ gives a marginal SCM $\mathcal{M}_{\text{marg}(\mathcal{L})}$ with a graph that is a subgraph of the latent projection of the graph of the SCM $\mathcal{M}$ onto $\mathcal{I} \setminus \mathcal{L}$. In general, not all marginalizations respect the latent projection, as is illustrated in the following example.

Example 2.5.10 (Marginalization does not respect the latent projection)**.** *Consider the SCM* $\mathcal{M}$ *of Example 2.3.11. Although* $\mathcal{M}$ *and its marginalization* $\mathcal{M}_{\text{marg}(\mathcal{L})}$ *with* $\mathcal{L} = \{2, 3\}$ *are interventionally equivalent w.r.t.* $\mathcal{I} \setminus \mathcal{L} = \{1, 4\}$*, the graph* $\mathcal{G}(\mathcal{M}_{\text{marg}(\mathcal{L})})$ *is not a subgraph of the latent projection of* $\mathcal{G}(\mathcal{M})$ *onto* $\mathcal{I} \setminus \mathcal{L}$*, as can be verified from the graphs depicted in Figure 2.4.*

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

Proposition 2.5.11. *Let* $\mathcal{M}$ *be an SCM that is ancestrally uniquely solvable w.r.t. a subset* $\mathcal{L} \subseteq \mathcal{I}$*. Then*

$$\big(\mathcal{G}^a \circ \text{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\text{marg}(\mathcal{L}) \circ \mathcal{G}^a\big)(\mathcal{M})$$

and

$$\big(\mathcal{G} \circ \text{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\text{marg}(\mathcal{L}) \circ \mathcal{G}\big)(\mathcal{M}).$$

The (augmented) graph of a marginalized SCM can be a strict subgraph of the corresponding latent projection if, for example, certain paths cancel each other out after the substitution of the measurable solution function(s) into the causal mechanism(s) on the margin (see Example 2.D.12). For acyclic SCMs, we recover with Proposition 2.5.11 the known result that this class is closed under marginalization (see Proposition 2.3.4) (Evans, 2016). For linear SCMs, we have that unique solvability w.r.t. a subset $\mathcal{L}$ holds if and only if ancestral unique solvability w.r.t. $\mathcal{L}$ holds (see Proposition 2.C.4), and hence, a marginalization of a linear SCM always respects the latent projection.

2.6 MARKOV PROPERTIES

In this section, we give a short overview of Markov properties for SCMs with cycles. We make use of the Markov properties that were recently developed by Forré and Mooij (2017) for HEDGes, a graphical representation that is similar to the

augmented graph of SCMs. We briefly summarize some of their main results and apply them to the class of SCMs. In Appendix 2.A.2, we provide a more thorough introduction and give an intuitive derivation, which can act as an entry point for the reader into the more extensive discussion of Markov properties provided in Forré and Mooij (2017).

Markov properties associate a set of conditional independence relations to a graph. The directed global Markov property for directed acyclic graphs (see Definitions 2.A.4 and 2.A.6), also known as the d-separation criterion (Pearl, 1985), is one of the most widely used. It directly extends to a similar property for acyclic directed mixed graphs (ADMGs) (Richardson, 2003). It does not hold in general for cyclic SCMs, however, as was already observed earlier (Spirtes, 1994, 1995).

Example 2.6.1 (Directed global Markov property does not hold for cyclic SCM)**.** *One can check that for every solution $\boldsymbol{X}$ of the SCM $\mathcal{M}$ of Example 2.3.5, X_1 is not independent of X_2 given $\{X_3, X_4\}$. However, the variables X_1 and X_2 are d-separated given $\{X_3, X_4\}$ in $\mathcal{G}(\mathcal{M})$ (see Figure 2.3). Hence the global directed Markov property does not hold here.*

Although some progress has been made in the case of discrete (Pearl and Dechter, 1996; Neal, 2000; Forré and Mooij, 2017) and linear models (Spirtes, 1993, 1994, 1995; Koster, 1996; Richardson, 1996c; Hyttinen, Eberhardt, and Hoyer, 2012; Forré and Mooij, 2017), only recently a general directed global Markov property has been introduced for more general cyclic models (Forré and Mooij, 2017), that is based on σ-separation (see Definition 2.A.16 and 2.A.20), an extension of d-separation. This notion of σ-separation was derived from the notion of d-separation in the acyclification of the graph (Forré and Mooij, 2017) (see Definition 2.A.13). The acyclification of a graph generalizes the idea of the collapsed graph developed by Spirtes (1994) and can, in particular, be applied to the graphs of SCMs. The main idea of the acyclification is that under the condition that the SCM is uniquely solvable w.r.t. each strongly connected component, we can replace the causal mechanisms of these strongly connected components by their measurable solution functions, which results in an acyclic SCM. This acyclified SCM (see Definition 2.A.11) is observationally equivalent to the original SCM (see Proposition 2.A.12).

Example 2.6.2 (Construction of an observationally equivalent acyclic SCM)**.** *The SCM $\mathcal{M}$ of Example 2.3.5 is uniquely solvable w.r.t. all its strongly connected components, that is, the subsets $\{1\}$, $\{2\}$ and $\{3, 4\}$. Replacing the causal mechanisms of these strongly connected components by their measurable solution functions gives the observationally equivalent SCM $\tilde{\mathcal{M}}$ of Example 2.4.2. Because $\tilde{\mathcal{M}}$ is acyclic (see Figure 2.3) we can apply the directed global Markov property to $\tilde{\mathcal{M}}$. The fact that X_1 and X_2 are not d-separated given $\{X_3, X_4\}$ in $\mathcal{G}(\tilde{\mathcal{M}})$ is in line with X_1 being dependent of X_2 given $\{X_3, X_4\}$ for every solution $\boldsymbol{X}$ of $\tilde{\mathcal{M}}$ (and hence of $\mathcal{M}$).*

This acyclification preserves solutions, and d-separation in the acyclification can directly be translated into σ-separation on the original graph (see Proposition 2.A.19). This leads to the general directed global Markov property. The following theorem summarizes the main results of (Forré and Mooij, 2017) applied to SCMs.

Theorem 2.6.3 (Global Markov properties for SCMs (Forré and Mooij, 2017))**.** *Let $\mathcal{M}$ be a uniquely solvable SCM. Then its observational distribution $\mathbb{P}^{\boldsymbol{X}}$ exists, is unique and the following two statements hold:*

1. $\mathbb{P}^{\boldsymbol{X}}$ *satisfies the* directed global Markov property *("d-separation criterion") relative to $\mathcal{G}(\mathcal{M})$ (see Definition 2.A.6) if $\mathcal{M}$ satisfies at least one of the following conditions:*
   a) *$\mathcal{M}$ is acyclic;*
   b) *all endogenous spaces $\mathcal{X}_i$ are discrete and $\mathcal{M}$ is ancestrally uniquely solvable;*
   c) *$\mathcal{M}$ is linear (see Definition 2.C.1), each of its causal mechanisms $\{f_i\}_{i\in\mathcal{I}}$ has a non-trivial dependence on at least one exogenous variable, and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{\mathcal{J}}$.*
2. $\mathbb{P}^{\boldsymbol{X}}$ *satisfies the* general directed global Markov property *("σ-separation criterion") relative to $\mathcal{G}(\mathcal{M})$ (see Definition 2.A.20) if $\mathcal{M}$ is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$.*[16]

The general directed global Markov property is generally weaker than the directed global Markov property, since σ-separation implies d-separation. The acyclic case is well known and was first shown in the context of linear-Gaussian structural equation models (Spirtes et al., 1998; Koster, 1999). The discrete case fixes the erroneous theorem by Pearl and Dechter (1996), for which a counterexample was found by Neal (2000), by adding the ancestral unique solvability condition, and extends it to allow for bidirected edges in the graph. The linear case is an extension of existing results for the linear-Gaussian setting without bidirected edges (Spirtes, 1994, 1995; Koster, 1996) to a linear (possibly non-Gaussian) setting with bidirected edges in the graph.

In constraint-based approaches to causal discovery, one usually assumes the converse of the (general) directed global Markov property to hold (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009), which is called σ-faithfulness respectively d-faithfulness (see Definition 2.A.9 and 2.A.23). Meek (1995) showed that for multinomial and linear-Gaussian DAG (i.e., acyclic and causally sufficient SCMs) models, d-faithfulness holds for all parameter values up to a measure zero set. Up to our knowledge no such results have been shown in more general parametric or non-parametric settings (neither for d-faitfhulness in acyclic or cyclic settings, nor for σ-faithfulness).

16 Since (Forré and Mooij, 2017) also provides results under the weaker condition that an SCM is solvable (not necessarily uniquely) w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$, one might believe that Theorem 2.6.3.(2) could be generalized to stating that in that case, any of its observational distributions satisfies the general directed global Markov property. However, that is not true: consider, for example, the SCM $\mathcal{M} = \langle \mathbf{2}, \emptyset, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ with $f_1(\boldsymbol{x}) = x_1$ and $f_2(\boldsymbol{x}) = x_2$. Then $\mathcal{M}$ is solvable w.r.t. each of its strongly connected components $\{1\}$ and $\{2\}$. The solution with $X_1 = X_2$, where X_2 has a nondegenerate distribution, shows a dependence between X_1 and X_2, and thus $X_1 \perp\!\!\!\perp X_2$ does not hold. In general, all strongly connected components that admit multiple solutions may be dependent on any other variable(s) in the model.

2.7 CAUSAL INTERPRETATION OF THE GRAPH OF SCMS

In Example 2.4.4, we already saw that sometimes no information in the observational, interventional and even the counterfactual distributions suffices to decide whether a directed path or bidirected edge is present in the graph, or not. Here, we do not attempt to provide a complete characterization of the conditions under which the presence or absence of a directed path or bidirected edge in the graph can be identified from the observational and interventional distributions. Instead, we give sufficient conditions to detect a directed path and bidirected edge in the graph.

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

2.7.1 *Directed paths and edges*

For cyclic SCMs, the causal interpretation of the SCM is not always consistent with its graph. This can be illustrated with the SCM $\mathcal{M}$ of Example 2.5.10. Here, one sees a difference in the marginal distribution $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{1\},\xi_1)}}$ on $\mathcal{X}_4$ for different values of ξ_1, although variable 1 is not an ancestor of variable 4 and each marginal distribution $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\{1\},\xi_1)}}$ on $\mathcal{X}_4$ is uniquely defined. This counterintuitive behavior that an intervention on a nonancestor of a variable can change the distribution of that variable was already observed by Neal (2000). However, under a specific unique solvability condition, we obtain a direct causal interpretation for the absence of a directed edge or directed path in the graph of an SCM.

Proposition 2.7.1 (Sufficient condition for detecting a directed edge in the latent projection of the graph of an SCM)**.** *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, a subset $\mathcal{O} \subseteq \mathcal{I}$ and $i, j \in \mathcal{O}$ such that $i \neq j$. Let $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, where $I := \mathcal{O} \setminus \{i, j\}$, such that $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is uniquely solvable w.r.t.* $\mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\setminus i}}(j)$*. If there exist values $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ such that both $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}$ and $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\tilde{\xi}_i)}$ induce unique marginal distributions on $\mathcal{X}_j$, and these two induced distributions do not coincide, that is, there exists a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that*

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

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

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

The condition in Proposition 2.7.1 is a sufficient condition for determining whether a directed edge or path is present in the graph. In general, not all directed edges and paths can be identified from the interventional distributions with

this sufficient condition. For example, no interventional distribution satisfies the condition of Proposition 2.7.1 for the SCM $\bar{\mathcal{M}}$ in Example 2.4.4, although there is a directed edge $1 \to 2$ in the graph $\mathcal{G}(\bar{\mathcal{M}})$.

2.7.2 *Bidirected edges*

It is well known that there exists a similar sufficient condition for detecting bidirected edges in the graph of an acyclic SCM also known as the common-cause principle (see, e.g., Pearl, 2009). In the two variables case, this criterion informally states that there exists a bidirected edge between the variables i and j in the graph of the SCM, if the marginal interventional distribution of X_j under the intervention $\mathrm{do}(\{i\}, x_i)$ differs from the conditional distribution of X_j given $X_i = x_i$ (see Example 2.D.13). The following proposition provides a generalization of this sufficient condition for detecting bidirected edges in graphs of SCMs that may include cycles.

Proposition 2.7.2 (Sufficient condition for detecting a bidirected edge in the latent projection of the graph of an SCM)**.** *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$, a subset $\mathcal{O} \subseteq \mathcal{I}$ and $i, j \in \mathcal{O}$ such that $i \neq j$. Let $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, where $I := \mathcal{O} \setminus \{i, j\}$, such that $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ is uniquely solvable w.r.t. both $\mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ and $\mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\setminus i}}(j)$. Assume that for every $\xi_i \in \mathcal{X}_i$ both $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$ and $(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})_{\mathrm{do}(\{i\},\xi_i)}$ induce a unique marginal distribution on $\mathcal{X}_j \times \mathcal{X}_i$ and $\mathcal{X}_j$, respectively. If $j \notin \mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ and there exists a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that for every version of the regular conditional probability $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B}_j \mid X_i = \xi_i)$, there exists a value $\xi_i \in \mathcal{X}_i$ such that*

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

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

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

2.8 SIMPLE SCMS

In this section, we introduce the well-behaved class of simple SCMs. Simple SCMs satisfy all the local unique solvability conditions to ensure that this class is closed

under both perfect intervention and marginalization. They extend the subclass of acyclic SCMs to the cyclic setting, while preserving many of their convenient properties.

Definition 2.8.1 (Simple SCM)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM. We call $\mathcal{M}$* simple *if it is uniquely solvable w.r.t. every subset $\mathcal{O} \subseteq \mathcal{I}$.*

Loosely speaking, an SCM is simple if any subset of its structural equations can be solved uniquely for its associated variables in terms of the other variables that appear in these equations. An example of a simple SCM is given in Example 2.D.1.

On simple SCMs one can perform any number of marginalizations (see Definition 2.5.3) in any order (see Proposition 2.5.4). All these marginalizations respect the latent projection (see Proposition 2.5.11) and each resulting marginal SCM is again simple. Moreover, we show that this class is closed under intervention and the twin operation.

Proposition 2.8.2. *The class of simple SCMs is closed under marginalization, perfect intervention and the twin operation.*

The class of simple SCMs contains the acyclic SCMs as a subclass (see Proposition 2.3.4). In particular, a simple SCM has no self-cycles (see Proposition 2.3.7), since a self-cycle denotes that that variable cannot be uniquely (up to a $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-null set) determined by its parents.

From Proposition 2.8.2, it follows that the results summarized in Theorem 2.6.3 also apply to all the observational, interventional and counterfactual distributions of simple SCMs.

Corollary 2.8.3 (Global Markov properties for simple SCMs)**.** *Let $\mathcal{M}$ be a simple SCM. Then the:*

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

all exist, are unique and satisfy the general directed global Markov property relative to $\mathcal{G}(\mathcal{M})$, $\mathrm{do}(I)(\mathcal{G}(\mathcal{M}))$ and $\mathrm{do}(\tilde{I})(\mathrm{twin}(\mathcal{G}(\mathcal{M})))$, respectively. Moreover, if $\mathcal{M}$ satisfies at least one of the three conditions (1a), (1b), (1c) of Theorem 2.6.3, then they also obey the directed global Markov property relative to $\mathcal{G}(\mathcal{M})$, $\mathrm{do}(I)(\mathcal{G}(\mathcal{M}))$ and $\mathrm{do}(\tilde{I})(\mathrm{twin}(\mathcal{G}(\mathcal{M})))$, respectively.

Many of these properties are also shown to hold for the class of *modular SCMs* (Forré and Mooij, 2017), which contains, in particular, the class of simple SCMs (see Appendix 2.A.3 for more details).

Moreover, simple SCMs satisfy the unique solvability conditions of Proposition 2.7.1 and 2.7.2, which allows us to define the causal relationships for simple SCMs in terms of its graph.

Definition 2.8.4 (Causal relationships for simple SCMs)**.** *Let $\mathcal{M}$ be a simple SCM.*

1. *If there exists a directed edge $i \to j \in \mathcal{G}(\mathcal{M})$, that is, $i \in \mathrm{pa}(j)$, then we call i* a direct cause of j according to $\mathcal{M}$;

2. *If there exists a directed path $i \to \cdots \to j$ in $\mathcal{G}(\mathcal{M})$, that is, $i \in \mathrm{an}(j)$, then we call i* a cause of j according to $\mathcal{M}$;

3. *If there exists a bidirected edge $i \leftrightarrow j \in \mathcal{G}(\mathcal{M})$, then we call i* and j (latently) confounded according to $\mathcal{M}$.

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

Corollary 2.8.5 (Sufficient conditions for the presence of causal and confoundedness relationships for simple SCMs)**.** *Let $\mathcal{M}$ be a simple SCM and $i, j \in \mathcal{I}$ such that $i \neq j$ and $I := \mathcal{I} \setminus \{i, j\}$. Then:*

1. *If there exist values $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ and $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ and a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that*

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

   then i is a direct cause of j according to $\mathcal{M}$, that is, $i \to j \in \mathcal{G}(\mathcal{M})$;

2. *If there exist values $\xi_i \neq \tilde{\xi}_i \in \mathcal{X}_i$ and a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that*

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

   then i is a cause of j according to $\mathcal{M}$, that is, $i \to \cdots \to j$ in $\mathcal{G}(\mathcal{M})$;

3. *If $j \notin \mathrm{an}_{\mathcal{G}(\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)})}(i)$ and there exist a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ and a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that for every version of the regular conditional probability $\mathbb{P}_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(X_j \in \mathcal{B}_j \mid X_i = \xi_i)$ there exists a value $\xi_i \in \mathcal{X}_i$ such that*

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

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

For simple SCMs, it is in general not possible to identify all the causal and confoundedness relationships in the graph from the observational, interventional or even the counterfactual distributions. Examples 2.4.4 and 2.D.10 show that this is already impossible for acyclic SCMs without further assumptions.

Finally, there is a connection between SCMs and potential outcomes (Rubin, 1974) that generalizes to the cyclic setting. One of the consequences of Proposition 2.8.2 is that all counterfactuals are defined for a simple SCM (even if it is cyclic). This allows us to define potential outcomes in terms of a simple SCM in the following way.

Definition 2.8.6 (Potential outcome)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be a simple SCM, $I \subseteq \mathcal{I}$ a subset, $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$ a value and $\boldsymbol{E}$ a random variable such that $\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$. The*

potential outcome under the perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *is defined as* $\boldsymbol{X}_{\boldsymbol{\xi}_I} := g_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}}(\boldsymbol{E}_{\mathrm{pa}(\mathcal{I})})$, *where* $g_{\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}} : \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{I})} \to \boldsymbol{\mathcal{X}}$ *is a measurable solution function for* $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$.

2.9 DISCUSSION

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

We introduced various notions of (unique) solvability and showed that under appropriate (unique) solvability conditions, many of the operations and results for the acyclic setting can be extended to SCMs with cycles. For example, we introduced several equivalence relations between SCMs to compare SCMs at different levels of abstraction, we showed how to define marginal SCMs on a subset of the variables that are (in various ways) equivalent to the original SCM, we discussed under which conditions the distributions satisfy the (general) directed global Markov property relative to their graphs and we showed under which conditions the graph of an SCM can be interpreted causally. Most of these results are shown under sufficient conditions that are not necessary (e.g., for the marginalization operation this was shown in Example 2.D.11). It may therefore be possible to further relax some of the conditions.

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

One key property of simple SCMs is that the solutions always satisfy the conditional independencies implied by σ-separation. By simply replacing d-separation with σ-separation it turns out that one can directly extend results and algorithms for acyclic SCMs to the more general class of simple SCMs. For example, adjustment criteria (including the back-door criterion), Pearl's do-calculus and Tian's ID algorithm for the identification of causal effects have been extended recently to the class of modular SCMs, which contains the class of simple SCMs (Forré and

Mooij, 2019). Several causal discovery algorithms have already been proposed that work with simple SCMs, for example, the first constraint-based causal discovery algorithm that can deal with cycles and nonlinear functional relationships (Forré and Mooij, 2018). Also, Local Causal Discovery (LCD) (Cooper, 1997), Y-structures (Mani, 2006) and the Joint Causal Inference framework (JCI) all apply to simple SCMs (Mooij, Magliacane, and Claassen, 2020) even though they were originally developed for acyclic SCMs only. Recently, it has been shown that even the well-known Fast Causal Inference (FCI) algorithm (Spirtes, Meek, and Richardson, 1999; Zhang, 2008) is directly applicable to simple SCMs (Mooij and Claassen, 2020) and provides a consistent estimate of the Markov equivalence class (under the faithfulness assumption). Moreover, a method for constructing nonlinear simple SCMs using neural networks and sampling from them has been proposed (Forré and Mooij, 2018). This illustrates that the class of simple SCMs forms a convenient and practical extension of the class of acyclic SCMs that can be used for the purposes of causal modeling, reasoning, discovery and prediction.

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

CHAPTER APPENDIX

These appendices to Chapter 2 contain a summary of the basic terminology and results for causal graphical models (Appendix 2.A), additional (unique) solvability properties (Appendix 2.B), some results for linear SCMs (Appendix 2.C), other examples (Appendix 2.D), the proofs of all the theoretical results (Appendix 2.E) and the measurable selection theorems (Appendix 2.F) that are used in several proofs.

2.A CAUSAL GRAPHICAL MODELS

In this appendix, we provide a summary of the basic terminology and results for causal graphical models. In Appendix 2.A.1 we provide the terminology for directed (mixed) graphs. In Appendix 2.A.2 we give an introduction and an intuitive derivation of Markov properties for SCMs with cycles. In Appendix 2.A.3 we provide a definition of modular SCMs and show how they relate to SCMs. In Appendix 2.A.4 we provide an overview of the causal graphical models related to SCMs. The proofs of the theoretical results in this appendix are given in Appendix 2.E.

2.A.1 *Directed (mixed) graphs*

In this subsection, we introduce the terminology for directed (mixed) graphs, where we do allow for cycles (Lauritzen, 1996; Richardson, 2003; Pearl, 2009; Forré and Mooij, 2017).

Definition 2.A.1 (Directed (mixed) graph)**.**

1. *A* directed graph *is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of directed edges, which is a subset $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ of ordered pairs of nodes. Each element $(i, j) \in \mathcal{E}$ can be represented by the directed edge $i \to j$ or equivalently $j \leftarrow i$. In particular, $(i, i) \in \mathcal{E}$ represents a* self-cycle *$i \to i$.*

2. *A* directed mixed graph *is a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$, where the pair $(\mathcal{V}, \mathcal{E})$ forms a directed graph and $\mathcal{B}$ is a set of bidirected edges, which is a subset $\mathcal{B} \subseteq \{\{i, j\} : i, j \in \mathcal{V}, i \neq j\}$ of unordered (distinct) pairs of nodes. Each element $\{i, j\} \in \mathcal{B}$ can be represented by the bidirected edge $i \leftrightarrow j$ or equivalently $j \leftrightarrow i$. Note that a directed graph can be considered as a directed mixed graph without bidirected edges.*

3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph. A directed mixed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$ is a* subgraph of *$\mathcal{G}$ if $\tilde{\mathcal{V}} \subseteq \mathcal{V}$, $\tilde{\mathcal{E}} \subseteq \mathcal{E}$ and $\tilde{\mathcal{B}} \subseteq \mathcal{B}$, in which case we write $\tilde{\mathcal{G}} \subseteq \mathcal{G}$. For a subset $\mathcal{W} \subseteq \mathcal{V}$, we define the* induced subgraph of $\mathcal{G}$ on $\mathcal{W}$ *by $\mathcal{G}_{\mathcal{W}} := (\mathcal{W}, \tilde{\mathcal{E}}, \tilde{\mathcal{B}})$, where $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{B}}$ are the set of directed and bidirected edges in $\mathcal{E}$ and $\mathcal{B}$, respectively, that lie in $\mathcal{W} \times \mathcal{W}$ and $\{\{i, j\} : i, j \in \mathcal{W}, i \neq j\}$, respectively.*

4. *A* walk *between $i, j \in \mathcal{V}$ in a directed mixed graph $\mathcal{G}$ is a tuple $(i_0, \epsilon_1, i_1, \epsilon_2, i_2, \ldots, \epsilon_n, i_n)$ of alternating nodes and edges in $\mathcal{G}$ for some $n \geq 0$, where all $i_0, \ldots, i_n \in \mathcal{V}$, all $\epsilon_1, \ldots, \epsilon_n \in \mathcal{E} \cup \mathcal{B}$ such that $\epsilon_k \in \{i_{k-1} \to i_k, i_{k-1} \leftarrow i_k, i_{k-1} \leftrightarrow i_k\}$ for all $k = 1, \ldots, n$, and it starts with node $i_0 = i$ and ends with node $i_n = j$. Note that $n = 0$ corresponds with a trivial walk consisting of a single node. If all nodes $i_0, \ldots, i_n$ are distinct, it is called a* path. *A walk (path) of the form $i \to \cdots \to j$, that is, ϵ_k is $i_{k-1} \to i_k$ for all $k = 1, 2, \ldots, n$, is called a* directed walk (path) *from i to j.*

5. *A* cycle *through $i \in \mathcal{V}$ in a directed mixed graph $\mathcal{G}$ is a directed path from i to some node j extended with the edge $j \to i \in \mathcal{E}$. In particular, a self-cycle $i \to i \in \mathcal{E}$ is a cycle. Note that a path cannot contain any cycles. A directed graph and a directed mixed graph are said to be* acyclic *if they contain no cycles, and are then referred to as a* directed acyclic graph (DAG) *and an* acyclic directed mixed graph (ADMG)*, respectively.*

6. *For a directed mixed graph $\mathcal{G}$ and a node $i \in \mathcal{V}$ we define the set of* parents *of i by $\mathrm{pa}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : j \to i \in \mathcal{E}\}$, the set of children of i by $\mathrm{ch}_{\mathcal{G}}(i) := \{j \in \mathcal{V} : i \to j \in \mathcal{E}\}$, the set of* ancestors *of i by*

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

and the set of descendants *of i by*

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

Note that we have $\{i\} \cup \mathrm{pa}_{\mathcal{G}}(i) \subseteq \mathrm{an}_{\mathcal{G}}(i)$ and $\{i\} \cup \mathrm{ch}_{\mathcal{G}}(i) \subseteq \mathrm{de}_{\mathcal{G}}(i)$. We can apply all these definitions to subsets $\mathcal{U} \subseteq \mathcal{V}$ by taking unions, for example $\mathrm{pa}_{\mathcal{G}}(\mathcal{U}) := \cup_{i \in \mathcal{U}} \mathrm{pa}_{\mathcal{G}}(i)$. A subset $\mathcal{A} \subseteq \mathcal{V}$ is called an ancestral subset *in $\mathcal{G}$ if $\mathcal{A} = \mathrm{an}_{\mathcal{G}}(\mathcal{A})$, that is, $\mathcal{A}$ is closed under taking ancestors of $\mathcal{A}$ in $\mathcal{G}$.*

7. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph. We call $\mathcal{G}$* strongly connected *if for every pair of distinct nodes $i, j \in \mathcal{V}$, the graph contains a cycle that passes through both i and j. The* strongly connected component of *$i \in \mathcal{V}$, denoted by $\mathrm{sc}_{\mathcal{G}}(i)$, is the maximal subset $\mathcal{S} \subseteq \mathcal{V}$ such that $i \in \mathcal{S}$ and the induced subgraph $\mathcal{G}_{\mathcal{S}}$ is strongly connected. Equivalently, $\mathrm{sc}_{\mathcal{G}}(i) = \mathrm{an}_{\mathcal{G}}(i) \cap \mathrm{de}_{\mathcal{G}}(i)$.*

8. *A* loop *in a directed mixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ is a subset $\mathcal{O} \subseteq \mathcal{V}$ that is strongly connected in the induced subgraph $\mathcal{G}_{\mathcal{O}}$ of $\mathcal{G}$ on $\mathcal{O}$.*

9. *For a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we define the* graph of strongly connected components *of $\mathcal{G}$ as the directed graph $\mathcal{G}^{\mathrm{sc}} := (\mathcal{V}^{\mathrm{sc}}, \mathcal{E}^{\mathrm{sc}})$, where $\mathcal{V}^{\mathrm{sc}}$ are the strongly connected components of $\mathcal{G}$, that is, $\mathcal{V}^{\mathrm{sc}}$ are the equivalence classes in $\mathcal{V}/\!\sim$ with the equivalence relation $i \sim j$ if and only if $i \in \mathrm{sc}_{\mathcal{G}}(j)$, and $\mathcal{E}^{\mathrm{sc}} = (\mathcal{E} \setminus \{i \to i : i \in \mathcal{V}\})/\!\sim$ with the equivalence relation $(i \to j) \sim (i' \to j')$ if and only if $i \sim i'$ and $j \sim j'$.*

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

Lemma 2.A.2 (DAG of strongly connected components)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph. Then $\mathcal{G}^{\text{sc}}$, the graph of strongly connected components of $\mathcal{G}$, is a DAG.*

2.A.2 *Markov properties*

In this subsection, we give a short overview of Markov properties for SCMs with cycles. We will make use of the Markov properties that were recently developed by Forré and Mooij (2017) for HEDGes, a graphical representation that is similar to the augmented graph of SCMs. We briefly summarize some of their main results and apply them to the class of SCMs. We also provide a shorter and more intuitive derivation so that this subsection can act as an entry point for the reader into the more extensive discussion of Markov properties provided in Forré and Mooij (2017).

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

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

For a more extensive study of different Markov properties that can be associated to SCMs we refer the reader to (Forré and Mooij, 2017).

2.A.2.1 *The directed global Markov property*

Conditional independencies in the observational distribution of an acyclic SCM can be read off from its graph by using the graphical criterion called d-separation (Pearl, 2009). The directed global Markov property associates a conditional independence relation in the observational distribution of the SCM to each d-separation entailed by the graph. Here, we use a formulation of d-separation that generalizes d-separation for DAGs (Pearl, 1985) and m-separation for ADMGs (Richardson, 2003) and mDAGs (Evans, 2016).

Definition 2.A.3 (Collider). *Let $\pi = (i_0, \epsilon_1, i_1, \epsilon_2, i_2, \ldots, \epsilon_n, i_n)$ be a walk (path) in a directed mixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$. A node i_k on π is called a* collider on π *if it is a non-endpoint node ($1 \leq k < n$) and the two edges $\epsilon_k, \epsilon_{k+1}$ meet head-to-head on i_k (i.e., if the subwalk $(i_{k-1}, \epsilon_k, i_k, \epsilon_{k+1}, i_{k+1})$ is of the form $i_{k-1} \to i_k \leftarrow i_{k+1}$, $i_{k-1} \leftrightarrow i_k \leftarrow i_{k+1}$, $i_{k-1} \to i_k \leftrightarrow i_{k+1}$ or $i_{k-1} \leftrightarrow i_k \leftrightarrow i_{k+1}$). The node i_k is called a* non-collider on π *otherwise, that is, if it is an endpoint node ($k = 0$ or $k = n$) or if the subwalk $(i_{k-1}, \epsilon_k, i_k, \epsilon_{k+1}, i_{k+1})$ is of the form $i_{k-1} \to i_k \to i_{k+1}$, $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$, $i_{k-1} \leftarrow i_k \to i_{k+1}$, $i_{k-1} \leftrightarrow i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftrightarrow i_{k+1}$.*

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

Definition 2.A.4 (d-separation). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and let $C \subseteq \mathcal{V}$ be a subset of nodes. A walk (path) $\pi = (i_0, \epsilon_1, i_1, \ldots, i_n)$ in $\mathcal{G}$ is said to be* C-d-blocked or d-blocked by C *if*

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

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

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

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

Lemma 2.A.5. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph, $C \subseteq \mathcal{V}$ and $i, j \in \mathcal{V}$. There exists a C-d-open walk between i and j in $\mathcal{G}$ if and only if there exists a C-d-open path between i and j in $\mathcal{G}$.*

Definition 2.A.6 (Directed global Markov property). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathbb{P}_{\mathcal{V}}$ a probability distribution on $\boldsymbol{\mathcal{X}}_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$, where each $\mathcal{X}_i$ is a standard probability space. The probability distribution $\mathbb{P}_{\mathcal{V}}$ satisfies the* directed global Markov property *relative to $\mathcal{G}$ if for all subsets $A, B, C \subseteq \mathcal{V}$ we have*

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

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

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

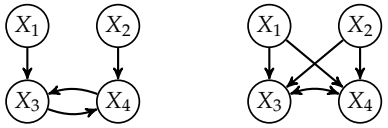


Figure 2.5: The graphs of the observationally equivalent SCMs $\mathcal{M}$ (left) and $\tilde{\mathcal{M}}$ (right) of Example 2.A.8 and 2.A.10.

Theorem 2.A.7 (Directed global Markov property for SCMs (Forré and Mooij, 2017))**.** *Let $\mathcal{M}$ be a uniquely solvable SCM that satisfies at least one of the following three conditions:*

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

Then its observational distribution $\mathbb{P}^{\boldsymbol{X}}$ exists, is unique and satisfies the directed global Markov property relative to $\mathcal{G}(\mathcal{M})$ (see Definition 2.A.6).

The acyclic case is well known and was first shown in the context of linear-Gaussian structural equation models (Spirtes et al., 1998; Koster, 1999). The discrete case fixes the erroneous theorem by Pearl and Dechter (1996), for which a counterexample was found by Neal (2000), by adding the ancestral unique solvability condition, and extends it to allow for bidirected edges in the graph. The linear case is an extension of existing results for the linear-Gaussian setting without bidirected edges (Spirtes, 1994, 1995; Koster, 1996) to a linear (possibly non-Gaussian) setting with bidirected edges in the graph.

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

Example 2.A.8 (Directed global Markov property does not hold for cyclic SCM)**.** *Consider the SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with $\mathcal{I} = \mathcal{J} = \mathbf{4}$, $\mathcal{X}_i = \mathcal{E}_i = (-1, 1)$ for $i = 1, 2$, and $\mathcal{X}_i = \mathcal{E}_i = \mathbb{R}$ for $i = 3, 4$, the causal mechanism given by*

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

and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ the standard-normal distribution on $\mathbb{R}^4$ restricted to $\boldsymbol{\mathcal{E}}$. The graph of $\mathcal{M}$ is depicted in Figure 2.5 on the left. The model is uniquely solvable (it is even simple). One can check that for every solution $\boldsymbol{X}$ of $\mathcal{M}$, X_1 is not independent of X_2 given $\{X_3, X_4\}$. However, the variables X_1 and X_2 are d-separated given $\{X_3, X_4\}$ in $\mathcal{G}(\mathcal{M})$. Hence the global directed Markov property does not hold here.

In constraint-based approaches to causal discovery, one usually assumes the converse of the directed global Markov property to hold (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009).

Definition 2.A.9 (d-Faithfulness)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathbb{P}_{\mathcal{V}}$ a probability distribution on $\boldsymbol{\mathcal{X}}_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$, where each $\mathcal{X}_i$ is a standard probability space. The probability distribution $\mathbb{P}_{\mathcal{V}}$ is d-*faithful *to $\mathcal{G}$ if for all subsets $A, B, C \subseteq \mathcal{V}$ we have*

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

where we take the canonical projections $X_i : \boldsymbol{\mathcal{X}}_{\mathcal{V}} \to \mathcal{X}_i$ as random variables.

In other words, the d-faithfulness assumption states that the graph explains, via d-separation, all the conditional independencies that are present in the observational distribution. Meek (1995) showed that for multinomial and linear-Gaussian DAG (i.e., acyclic and causally sufficient SCMs) models, d-faithfulness holds for all parameter values up to a measure zero set (in a natural parameterization). Up to our knowledge no such results have been shown in more general parametric or nonparametric settings (neither in the acyclic case, nor in the cyclic one).

2.A.2.2 *The general directed global Markov property*

In (Forré and Mooij, 2017) the general directed global Markov property is introduced, that is based on σ-separation, an extension of d-separation. This notion of σ-separation was derived from the notion of d-separation in the acyclification of the graph. The acyclification of a graph generalizes the idea of the collapsed graph for directed graphs, developed by Spirtes (1994), to HEDGes. In particular, this notion can be applied to directed mixed graphs, and thus to the graphs of SCMs. The main idea of the acyclification is that under the condition that the SCM is uniquely solvable w.r.t. each strongly connected component, we can replace the causal mechanisms of these strongly connected components by their measurable solution functions, which results in an acyclic SCM. This acyclification preserves the solutions, and d-separation in the acyclification can directly be translated into σ-separation in the original graph. This then leads to the general directed global Markov property. We will discuss this now in more detail.

Example 2.A.10 (Construction of an observationally equivalent acyclic SCM)**.** *Consider the SCM $\mathcal{M}$ of Example 2.A.8 which is uniquely solvable w.r.t. all its strongly connected components, i.e., the subsets $\{1\}$, $\{2\}$ and $\{3,4\}$. Replacing the causal mechanisms of these strongly connected components by their measurable solution functions gives the SCM $\tilde{\mathcal{M}}$ that is the same as $\mathcal{M}$ except that its causal mechanism $\tilde{\boldsymbol{f}}$ is given by*

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

By construction, $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are observationally equivalent. Because $\tilde{\mathcal{M}}$ is acyclic (see Figure 2.5 on the right) we can apply the directed global Markov property to $\tilde{\mathcal{M}}$. The fact that X_1 and X_2 are not d-separated given $\{X_3, X_4\}$ in $\mathcal{G}(\tilde{\mathcal{M}})$ is in line with X_1 being dependent of X_2 given $\{X_3, X_4\}$ for every solution $\boldsymbol{X}$ of $\tilde{\mathcal{M}}$ (and hence of $\mathcal{M}$).

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

Definition 2.A.11 (Acyclification of an SCM)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. For each $i \in \mathcal{I}$, let g_i be the i^{th} component of a measurable solution function $\boldsymbol{g}_{\mathrm{sc}(i)} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathrm{sc}(i))\setminus\mathrm{sc}(i)} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathrm{sc}(i))} \to \boldsymbol{\mathcal{X}}_{\mathrm{sc}(i)}$ of $\mathcal{M}$ w.r.t.* $\mathrm{sc}(i)$*, where* pa *and* sc *denote the parents and strongly connected components according to $\mathcal{G}^a(\mathcal{M})$, respectively. We call the SCM $\mathcal{M}^{\mathrm{acy}} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \hat{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with the* acyclified causal mechanism $\hat{\boldsymbol{f}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ *given by*

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

an acyclification of $\mathcal{M}$*. We denote by* $\mathrm{acy}(\mathcal{M})$ *the equivalence class of the acyclifications of $\mathcal{M}$.*

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

Proposition 2.A.12. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. Then an acyclification $\mathcal{M}^{\mathrm{acy}}$ of $\mathcal{M}$ is acyclic and observationally equivalent to $\mathcal{M}$.*

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

Definition 2.A.13 (Acyclification of a directed mixed graph)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph. The* acyclification of $\mathcal{G}$ *maps $\mathcal{G}$ to the* acyclified graph $\mathcal{G}^{\mathrm{acy}} := (\mathcal{V}, \hat{\mathcal{E}}, \hat{\mathcal{B}})$ *with directed edges $j \to i \in \hat{\mathcal{E}}$ if and only if $j \in \mathrm{pa}_{\mathcal{G}}(\mathrm{sc}_{\mathcal{G}}(i)) \setminus \mathrm{sc}_{\mathcal{G}}(i)$ and bidirected edges $i \leftrightarrow j \in \hat{\mathcal{B}}$ if and only if there exist $i' \in \mathrm{sc}_{\mathcal{G}}(i)$ and $j' \in \mathrm{sc}_{\mathcal{G}}(j)$ with $i' = j'$ or $i' \leftrightarrow j' \in \mathcal{B}$.*

The following compatibility result is immediate from the definitions.

Proposition 2.A.14. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. Then $\mathcal{G}^a(\mathrm{acy}(\mathcal{M})) \subseteq \mathrm{acy}(\mathcal{G}^a(\mathcal{M}))$ and $\mathcal{G}(\mathrm{acy}(\mathcal{M})) \subseteq \mathrm{acy}(\mathcal{G}(\mathcal{M}))$.*

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

Example 2.A.15 (Graph of the acyclification of the SCM is a strict subgraph of the acyclification of its graph)**.** *Consider the SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ with the causal mechanism defined by*

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

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

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

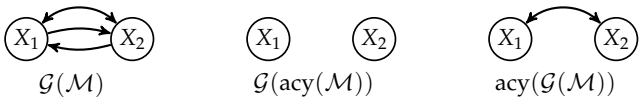


Figure 2.6: The graphs of the original SCM $\mathcal{M}$ (left), of the acyclified SCM (center), and of the acyclification of the graph of $\mathcal{M}$ (right) corresponding to Example 2.A.15.

The graph $\mathcal{G}(\text{acy}(\mathcal{M}))$ *is a strict subgraph of* $\text{acy}(\mathcal{G}(\mathcal{M}))$ *as can be seen in Figure 2.6.*

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

Definition 2.A.16 (σ-separation (Forré and Mooij, 2017))**.** *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *be a directed mixed graph and let* $C \subseteq \mathcal{V}$ *be a subset of nodes. A walk (path)* $\pi = (i_0, \epsilon_1, i_1, \ldots, i_n)$ *in* $\mathcal{G}$ *is said to be* C-σ-blocked or σ-blocked by C *if*

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

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

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

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

Although for proofs it is often easier to make use of walks, it suffices to formulate σ-separation in term of paths rather than walks because of the following result, which is analogous to a similar result for *d*-separation (see Lemma 2.A.5).

Lemma 2.A.17. *Let* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ *be a directed mixed graph,* $C \subseteq \mathcal{V}$ *and* $i, j \in \mathcal{V}$*. There exists a* C-σ*-open walk between* i *and* j *in* $\mathcal{G}$ *if and only if there exists a* C-σ*-open path between* i *and* j *in* $\mathcal{G}$*.*

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

Example 2.A.18 (*d*-separation does not imply σ-separation)**.** *Consider the directed graph* $\mathcal{G}$ *as depicted in Figure 2.5 (left). Here* X_1 *is d-separated from* X_2 *given* $\{X_3, X_4\}$*, but* X_1 *is not* σ*-separated from* X_2 *given* $\{X_3, X_4\}$*.*

The following result in (Forré and Mooij, 2017) relates σ-separation to d-separation.

Proposition 2.A.19. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph. Then for $A, B, C \subseteq \mathcal{V}$,*

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

By replacing in Definition 2.A.6 "d-separation" by "σ-separation", one obtains the formulation of what Forré and Mooij (2017) termed the general directed global Markov property.

Definition 2.A.20 (General directed global Markov property (Forré and Mooij, 2017))**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathbb{P}_{\mathcal{V}}$ a probability distribution on $\boldsymbol{\mathcal{X}}_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$, where each $\mathcal{X}_i$ is a standard probability space. The probability distribution $\mathbb{P}_{\mathcal{V}}$ satisfies the* general directed global Markov property *relative to $\mathcal{G}$ if for all subsets $A, B, C \subseteq \mathcal{V}$ we have*

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

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

The fact that σ-separation implies d-separation means that the directed global Markov property implies the general directed global Markov property. In other words, the general directed global Markov property is weaker than the directed global Markov property. It is actually strictly weaker, as we saw in Example 2.A.18.

The following fundamental result, also known as the σ-separation criterion, follows directly from the theory in (Forré and Mooij, 2017).

Theorem 2.A.21 (General directed global Markov property for SCMs)**.** *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$. Then its observational distribution $\mathbb{P}^{\boldsymbol{X}}$ exists, is unique and it satisfies the general directed global Markov property relative to $\mathcal{G}(\mathcal{M})$.*[17]

The proof is based on the reasoning that, for $A, B, C \subseteq \mathcal{I}$, if A is σ-separated from B given C in $\mathcal{G}(\mathcal{M})$, then A is d-separated from B by C in $\mathrm{acy}(\mathcal{G}(\mathcal{M}))$ and hence in $\mathcal{G}(\mathrm{acy}(\mathcal{M}))$, and since $\mathrm{acy}(\mathcal{M})$ is acyclic and observationally equivalent to $\mathcal{M}$, it follows from the directed global Markov property applied to $\mathrm{acy}(\mathcal{M})$ that $\boldsymbol{X}_A \perp\!\!\!\perp_{\mathbb{P}^{\boldsymbol{X}}} \boldsymbol{X}_B \,|\, \boldsymbol{X}_C$ for every solution $\boldsymbol{X}$ of $\mathcal{M}$. Note that the ancestral unique solvability condition for the discrete case is strictly weaker than the condition of unique solvability w.r.t. each strongly connected component in Theorem 2.A.21. For

17 Since (Forré and Mooij, 2017) also provides results under the weaker condition that an SCM is solvable (not necessarily uniquely) w.r.t. each strongly connected component of $\mathcal{G}(\mathcal{M})$, one might believe that Theorem 2.A.21 could be generalized to stating that in that case, any of its observational distributions satisfies the general directed global Markov property. However, that is not true: consider for example the SCM $\mathcal{M} = \langle \mathbf{2}, \emptyset, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ with $f_1(\boldsymbol{x}) = x_1$ and $f_2(\boldsymbol{x}) = x_2$. Then $\mathcal{M}$ is solvable w.r.t. each of its strongly connected components $\{1\}$ and $\{2\}$. The solution with $X_1 = X_2$ shows a dependence between X_1 and X_2 and thus $X_1 \perp\!\!\!\perp X_2$ does not hold. In general, all strongly connected components that admit multiple solutions may be dependent on any other variable(s) in the model.

the linear case, the condition of unique solvability is equivalent to the condition of unique solvability w.r.t. each strongly connected component (see Proposition 2.C.4).

The results in Theorems 2.A.7 and 2.A.21 are not preserved under perfect intervention, because intervening on a strongly connected component could split it into several strongly connected components with different solvability properties. As the class of simple SCMs is preserved under perfect intervention and the twin operation (Proposition 2.8.2), we obtain the following corollary.

Corollary 2.A.22 (Global Markov properties for simple SCMs)**.** *Let $\mathcal{M}$ be a simple SCM. Then the:*

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

all exist, are unique and satisfy the general directed global Markov property relative to $\mathcal{G}(\mathcal{M})$, $\mathrm{do}(I)(\mathcal{G}(\mathcal{M}))$ and $\mathrm{do}(\tilde{I})(\mathrm{twin}(\mathcal{G}(\mathcal{M})))$, respectively. Moreover, if $\mathcal{M}$ satisfies at least one of the three conditions (1), (2), (3) of Theorem 2.A.7, then they also satisfies the directed global Markov property relative to $\mathcal{G}(\mathcal{M})$, $\mathrm{do}(I)(\mathcal{G}(\mathcal{M}))$ and $\mathrm{do}(\tilde{I})(\mathrm{twin}(\mathcal{G}(\mathcal{M})))$, respectively.

Similar to *d*-faithfulness, σ-faithfulness[18] is defined as follows.

Definition 2.A.23 (σ-Faithfulness)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B})$ be a directed mixed graph and $\mathbb{P}_{\mathcal{V}}$ a probability distribution on $\boldsymbol{\mathcal{X}}_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \mathcal{X}_i$, where each $\mathcal{X}_i$ is a standard probability space. The probability distribution $\mathbb{P}_{\mathcal{V}}$ is* σ-faithful *to $\mathcal{G}$ if for all subsets $A, B, C \subseteq \mathcal{V}$ we have*

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

where we take the canonical projections $X_i : \boldsymbol{\mathcal{X}}_{\mathcal{V}} \to \mathcal{X}_i$ as random variables.

In other words, the graph explains, via σ-separation, all the conditional independencies that are present in the observational distribution. Although it has been conjectured (Spirtes, 1995) that under certain conditions σ-faithfulness should hold, formulating and proving such completeness results is an open problem to the best of our knowledge.

2.A.3 *Modular SCMs*

In this subsection, we relate the class of (simple) SCMs to that of modular SCMs. Modular SCMs introduced by Forré and Mooij (2017) are causal graphical models on which marginalizations and interventions are defined and they satisfy the general directed global Markov property. For a comprehensive account on modular SCMs we refer the reader to (Forré and Mooij, 2017).

18 In (Richardson, 1996c) it is called "collapsed graph faithfulness".

2.A.3.1 *Definition of a modular SCM*

In contrast to an SCM from which a graph can be derived, a modular SCM is defined in terms of a graphical object, which Forré and Mooij (2017) call a directed graph with hyperedges (HEDG). The hyperedges of a HEDG are described in terms of a simplicial complex.

Definition 2.A.24 (Simplicial complex)**.** *Let $\mathcal{V}$ be a finite set. A* simplicial complex $\mathcal{H}$ *over $\mathcal{V}$ is a set of subsets of $\mathcal{V}$ such that*

1. *all single element sets $\{v\}$ are in $\mathcal{H}$ for $v \in \mathcal{V}$, and*
2. *if $\mathcal{F} \in \mathcal{H}$, then also all subsets $\tilde{\mathcal{F}} \subseteq F$ are elements of $\mathcal{H}$.*

Definition 2.A.25 (Directed graph with hyperedges (HEDGes) (Forré and Mooij, 2017))**.** *A* directed graph with hyperedges (HEDG) *is a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$, where $(\mathcal{V}, \mathcal{E})$ is a directed graph and $\mathcal{H}$ a simplicial complex over the set of nodes $\mathcal{V}$. The elements $\mathcal{F}$ of $\mathcal{H}$ are called* hyperedges *of $\mathcal{G}$. The elements $\mathcal{F}$ of $\mathcal{H}$ that are inclusion-maximal elements of $\mathcal{H}$ are called* maximal hyperedges *and are denoted by $\hat{\mathcal{H}}$.*

A HEDG $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ can be represented as a directed graph $\bar{\mathcal{G}} := (\mathcal{V}, \mathcal{E})$ consisting of nodes $\mathcal{V}$ and directed edges $\mathcal{E}$, with additional maximal hyperedges $\mathcal{F} \in \hat{\mathcal{H}}$ with $|\mathcal{F}| \geq 2$ (i.e., not corresponding to single element sets $\{v\} \in \hat{\mathcal{H}}$), that point to their target nodes $v \in \mathcal{F}$. For a HEDG $\mathcal{G}$, we define $\mathrm{pa}_{\mathcal{G}}$, $\mathrm{ch}_{\mathcal{G}}$, etc., in terms of the underlying directed graph $\bar{\mathcal{G}}$, that is, $\mathrm{pa}_{\bar{\mathcal{G}}}$, $\mathrm{ch}_{\bar{\mathcal{G}}}$, etc., respectively.

A *loop* in a HEDG $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ is a subset $\mathcal{O} \subseteq \mathcal{V}$ that is a loop in the underlying directed graph $\bar{\mathcal{G}} = (\mathcal{V}, \mathcal{E})$. In other words, a loop of $\mathcal{G}$ is a set of nodes $\mathcal{O} \subseteq \mathcal{V}$ such that for every two nodes $v, w \in \mathcal{O}$ there are directed paths $v \to \cdots \to w$ and $w \to \cdots \to v$ in $\mathcal{G}$ for which all the intermediate nodes lie in $\mathcal{O}$ (if any exist). In particular, a loop may consist of a single element $\{v\}$ for $v \in \mathcal{V}$. The set of loops in $\mathcal{G}$ is denoted by $\mathcal{L}(\mathcal{G})$.

In order to define a modular SCM one needs the notion of a compatible system of solution functions, which assigns to each loop a separate solution function such that all these solution functions are "compatible" with each other.

Definition 2.A.26 (Compatible system of solution functions[19])**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ be a HEDG. For every $v \in \mathcal{V}$ and maximal hyperedge $\mathcal{F}$ in $\hat{\mathcal{H}}$, let $\mathcal{X}_v$ and $\mathcal{E}_{\mathcal{F}}$ be standard measurable spaces. For a subset $\mathcal{O} \subseteq \mathcal{V}$ we define*[20]

$$\boldsymbol{\mathcal{X}}_{\mathcal{O}} := \prod_{v \in \mathcal{O}} \mathcal{X}_v \quad \text{and} \quad \widehat{\boldsymbol{\mathcal{E}}}_{\mathcal{O}} := \prod_{\substack{\mathcal{F} \in \hat{\mathcal{H}} \\ \mathcal{F} \cap \mathcal{O} \neq \emptyset}} \mathcal{E}_{\mathcal{F}}\,.$$

Consider a family of measurable mappings $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O} \in \mathcal{L}(\mathcal{G})}$ indexed by $\mathcal{L}(\mathcal{G})$ which are of the form

$$\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}_{\mathcal{G}}(\mathcal{O}) \setminus \mathcal{O}} \times \widehat{\boldsymbol{\mathcal{E}}}_{\mathcal{O}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}\,.$$

19 We deviate from the terminology in (Forré and Mooij, 2017) where this is called a "compatible system of structural equations".

20 We use the "hat" notation $\widehat{\boldsymbol{\mathcal{E}}}_{\mathcal{O}}$ to distinguish it from the ordinary subscript convention that $\boldsymbol{\mathcal{E}}_{\mathcal{O}} = \prod_{\mathcal{F} \in \mathcal{O}} \mathcal{E}_{\mathcal{F}}$ for some subset $\mathcal{O} \subseteq \hat{\mathcal{H}}$.

We call the family of measurable mappings $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}$ *a* compatible system of solution functions, *if for all* $\mathcal{O}, \tilde{\mathcal{O}} \in \mathcal{L}(\mathcal{G})$ *with* $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ *and for all* $\widehat{\boldsymbol{e}}_{\mathcal{O}} \in \widehat{\boldsymbol{\mathcal{E}}}_{\mathcal{O}}$ *and* $\boldsymbol{x}_{\mathrm{pa}_{\mathcal{G}}(\mathcal{O})\cup\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}_{\mathcal{G}}(\mathcal{O})\cup\mathcal{O}}$ *we have*

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

This structure of a compatible system of solution functions is at the heart of the definition of a modular SCM.

Definition 2.A.27 (Modular structural causal model (mSCM) (Forré and Mooij, 2017))**.** *A* modular structural causal model (mSCM) *is a tuple*

$$\widehat{\mathcal{M}} := \langle \mathcal{G}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle\,,$$

where

1. $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ *is a HEDG,*
2. $\boldsymbol{\mathcal{X}} = \prod_{v\in\mathcal{V}} \mathcal{X}_v$ *is the product of standard measurable spaces* $\mathcal{X}_v$*,*
3. $\boldsymbol{\mathcal{E}} = \prod_{\mathcal{F}\in\hat{\mathcal{H}}} \mathcal{E}_{\mathcal{F}}$ *is the product of standard measurable spaces* $\mathcal{E}_{\mathcal{F}}$*,*
4. $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}$ *is a compatible system of solution functions,*
5. $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{\mathcal{F}\in\hat{\mathcal{H}}} \mathbb{P}_{\mathcal{E}_{\mathcal{F}}}$ *is a product measure, where* $\mathbb{P}_{\mathcal{E}_{\mathcal{F}}}$ *is a probability measure on* $\mathcal{E}_{\mathcal{F}}$ *for each* $\mathcal{F} \in \hat{\mathcal{H}}$*.*

Let $\widehat{\mathcal{M}} = \langle \mathcal{G}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be a modular SCM and $\mathcal{O}_1, \ldots, \mathcal{O}_r \in \mathcal{L}(\mathcal{G})$ the strongly connected components of $\mathcal{G}$ ordered according to a topological order of the DAG of strongly connected components of $\mathcal{G}$. Then for any random variable $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$ such that $\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$ one can inductively define the random variables $X_v := (\boldsymbol{g}_{\mathcal{O}_i})_v(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{G}}(\mathcal{O}_i)\setminus\mathcal{O}_i}, \widehat{\boldsymbol{E}}_{\mathcal{O}_i})$ for all $v \in \mathcal{O}_i$ for all $i \geq 1$, starting at $X_v := (\boldsymbol{g}_{\mathcal{O}_1})_v(\widehat{\boldsymbol{E}}_{\mathcal{O}_1})$ for all $v \in \mathcal{O}_1$. Because $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}$ is a compatible system of solution functions, we have for every $\mathcal{O} \in \mathcal{L}(\mathcal{G})$

$$\boldsymbol{X}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{X}_{\mathrm{pa}_{\mathcal{G}}(\mathcal{O})\setminus\mathcal{O}}, \widehat{\boldsymbol{E}}_{\mathcal{O}})\,.$$

We call the random variable $\boldsymbol{X}$ a *solution* of the modular SCM $\widehat{\mathcal{M}}$. Note that the solution $\boldsymbol{X}$ depends on the choice of the random variable $\boldsymbol{E} : \Omega \to \boldsymbol{\mathcal{E}}$.

The causal semantics of modular SCMs can be defined in terms of perfect interventions, which is defined as follows.

Definition 2.A.28 (Perfect intervention on an mSCM)**.** *Consider a modular SCM* $\widehat{\mathcal{M}} = \langle \mathcal{G}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$*, a subset* $I \subseteq \mathcal{V}$ *of endogenous variables and a value* $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$*. The* perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$ *maps* $\widehat{\mathcal{M}}$ *to the modular SCM*

$$\widehat{\mathcal{M}}_{\mathrm{do}(I,\boldsymbol{\xi}_I)} := \langle \mathcal{G}^{\mathrm{do}}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}^{\mathrm{do}}, (\boldsymbol{g}^{\mathrm{do}}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G}^{\mathrm{do}})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}^{\mathrm{do}}} \rangle\,,$$

where

1. $\mathcal{G}^{\mathrm{do}} = (\mathcal{V}, \mathcal{E}^{\mathrm{do}}, \mathcal{H}^{\mathrm{do}})$, *where*

$$\mathcal{E}^{\mathrm{do}} = \mathcal{E} \setminus \{v \to w : v \in \mathcal{V}, w \in I\}$$

$$\mathcal{H}^{\mathrm{do}} = \{\mathcal{F} \setminus I : \mathcal{F} \in \mathcal{H}\} \cup \{\{v\} : v \in I\},$$

2. $\phi : \{\mathcal{F} \in \hat{\mathcal{H}} : \mathcal{F} \setminus I \neq \emptyset\} \to \hat{\mathcal{H}}^{\mathrm{do}} \setminus \{\{v\} : v \in I\}$ *is a mapping such that* $\phi(\mathcal{F}) \supseteq \mathcal{F} \setminus I$ *for all* $\mathcal{F} \in \hat{\mathcal{H}}$ *for which* $\mathcal{F} \setminus I \neq \emptyset$,

3. $\boldsymbol{\mathcal{E}}^{\mathrm{do}} = \prod_{\tilde{\mathcal{F}} \in \hat{\mathcal{H}}^{\mathrm{do}}} \mathcal{E}^{\mathrm{do}}_{\tilde{\mathcal{F}}}$, *where*

$$\mathcal{E}^{\mathrm{do}}_{\tilde{\mathcal{F}}} = \begin{cases} \mathcal{X}_v & \text{if } \tilde{\mathcal{F}} = \{v\} \text{ for } v \in I \\ \prod_{\mathcal{F} = \phi^{-1}(\tilde{\mathcal{F}})} \mathcal{E}_{\mathcal{F}} & \text{if } \tilde{\mathcal{F}} \in \hat{\mathcal{H}}^{\mathrm{do}} \setminus \{\{v\} : v \in I\}, \end{cases}$$

4. *for every* $\mathcal{O} \in \mathcal{L}(\mathcal{G}^{\mathrm{do}})$

$$\boldsymbol{g}^{\mathrm{do}}_{\mathcal{O}} = \begin{cases} \mathbb{I}_{\{v\}} & \text{if } \mathcal{O} = \{v\} \text{ for } v \in I \\ \boldsymbol{g}_{\mathcal{O}} & \text{otherwise,} \end{cases}$$

(note that if $\mathcal{O}$ *is a loop in* $\mathcal{G}^{\mathrm{do}}$*, then it is a loop in* $\mathcal{G}$*),*

5. $\mathbb{P}_{\boldsymbol{\mathcal{E}}^{\mathrm{do}}} = \prod_{\tilde{\mathcal{F}} \in \hat{\mathcal{H}}^{\mathrm{do}}} \mathbb{P}_{\mathcal{E}^{\mathrm{do}}_{\tilde{\mathcal{F}}}}$, *where*

$$\mathbb{P}_{\mathcal{E}^{\mathrm{do}}_{\tilde{\mathcal{F}}}} = \begin{cases} \delta_{\xi_v} & \text{if } \tilde{\mathcal{F}} = \{v\} \text{ for } v \in I \\ \prod_{\mathcal{F} = \phi^{-1}(\tilde{\mathcal{F}})} \mathbb{P}_{\mathcal{E}_{\mathcal{F}}} & \text{if } \tilde{\mathcal{F}} \in \hat{\mathcal{H}}^{\mathrm{do}} \setminus \{\{v\} : v \in I\}. \end{cases}$$

In contrast to SCMs, these perfect interventions on modular SCMs are directly defined on the underlying HEDG and depend on the choice of the mapping ϕ.

2.A.3.2 *Relation between SCMs and modular SCMs*

The solutions of a modular SCM can be described by an SCM that is loop-wisely solvable.

Definition 2.A.29 (Underlying SCM)**.** *Let* $\widehat{\mathcal{M}} = \langle \mathcal{G}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O} \in \mathcal{L}(\mathcal{G})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be a modular SCM. Then the mapping* ι *maps* $\widehat{\mathcal{M}}$ *to the* underlying SCM $\tilde{\mathcal{M}} := \langle \tilde{\mathcal{I}}, \tilde{\mathcal{J}}, \tilde{\boldsymbol{\mathcal{X}}}, \tilde{\boldsymbol{\mathcal{E}}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} \rangle$*, where*

1. $\tilde{\mathcal{I}} = \mathcal{V}$,
2. $\tilde{\mathcal{J}} = \hat{\mathcal{H}}$,
3. $\tilde{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}}$,
4. $\tilde{\boldsymbol{\mathcal{E}}} = \boldsymbol{\mathcal{E}}$,
5. $\tilde{\boldsymbol{f}}$ *is given by* $\tilde{f}_v = (\boldsymbol{g}_{\{v\}})_v$ *for all* $v \in \mathcal{V}$,

6. $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$.

Every solution $\boldsymbol{X}$ of a modular SCM $\widehat{\mathcal{M}}$ is also a solution of the underlying SCM $\iota(\widehat{\mathcal{M}})$.

Observe that for the modular SCM $\widehat{\mathcal{M}}$ we have that the induced subgraph $\mathcal{G}^a(\iota(\widehat{\mathcal{M}}))_{\tilde{\mathcal{I}}}$, of the augmented graph of the underlying SCM $\mathcal{G}^a(\iota(\widehat{\mathcal{M}}))$ on $\tilde{\mathcal{I}}$, is a subgraph of the underlying HEDG $\mathcal{G}$, that is, $\mathcal{G}^a(\iota(\widehat{\mathcal{M}}))_{\tilde{\mathcal{I}}} \subseteq \mathcal{G}$. This implies that, in general, the underlying HEDG $\mathcal{G}$ of $\widehat{\mathcal{M}}$ may have more loops than the loops in $\mathcal{G}(\iota(\widehat{\mathcal{M}}))$. For a subset $\mathcal{O} \subseteq \tilde{\mathcal{I}}$, we have for the exogenous parents of the underlying SCM $\iota(\widehat{\mathcal{M}})$

$$\mathrm{pa}(\mathcal{O}) \cap \tilde{\mathcal{J}} \subseteq \{\mathcal{F} \in \tilde{\mathcal{J}} \,:\, \mathcal{F} \cap \mathcal{O} \neq \emptyset\}\,,$$

where $\mathrm{pa}(\mathcal{O})$ denotes the set of parents of $\mathcal{O}$ in $\mathcal{G}^a(\iota(\widehat{\mathcal{M}}))$. Hence, in general, not all the hyperedges $\mathcal{F} \in \mathcal{H}$ such that $|\mathcal{F}| = 2$ (i.e., bidirected edges) are in the set of bidirected edges $\mathcal{B}$ of the graph of the underlying SCM $\mathcal{G}(\iota(\widehat{\mathcal{M}})) = (\mathcal{V}, \mathcal{E}, \mathcal{B})$. We conclude that the graph of the underlying SCM is, in general, a sparser graph than the HEDG of the modular SCM.

Next, we show that the compatible system of solution functions of a modular SCM induces a compatible system of solution functions on the underlying SCM. For this we need the notion of loop-wise solvability for SCMs.

Definition 2.A.30 (Loop-wise (unique) solvability for SCMs)**.** *We call an SCM $\mathcal{M}$*

1. loop-wisely solvable*, if $\mathcal{M}$ is solvable w.r.t. every loop $\mathcal{O} \in \mathcal{L}(\mathcal{G}(\mathcal{M}))$, and*
2. loop-wisely uniquely solvable*, if $\mathcal{M}$ is uniquely solvable w.r.t. every loop $\mathcal{O} \in \mathcal{L}(\mathcal{G}(\mathcal{M}))$.*

Definition 2.A.31 (Compatible system of solution functions for SCMs)**.** *For a loop-wisely solvable SCM $\mathcal{M}$, we call a family of measurable solution functions $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O} \in \mathcal{L}(\mathcal{G}(\mathcal{M}))}$, where $\boldsymbol{g}_{\mathcal{O}}$ is a measurable solution function of $\mathcal{M}$ w.r.t. $\mathcal{O}$, a* compatible system of solution functions*, if for all $\mathcal{O}, \tilde{\mathcal{O}} \in \mathcal{L}(\mathcal{G}(\mathcal{M}))$ with $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ and for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ we have*

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

The underlying SCM of a modular SCM always has a compatible system of solution functions, by construction.

Proposition 2.A.32. *Let $\widehat{\mathcal{M}} = \langle \mathcal{G}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, (\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O} \in \mathcal{L}(\mathcal{G})}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be a modular SCM. Then the underlying SCM $\tilde{\mathcal{M}} := \iota(\widehat{\mathcal{M}})$ is loop-wisely solvable. Moreover, it has a compatible system of solution functions $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O} \in \mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))}$, where $\boldsymbol{g}_{\mathcal{O}}$ is a measurable solution function of $\tilde{\mathcal{M}}$ w.r.t. $\mathcal{O}$.*

This shows that a modular SCM can be seen as an SCM together with an additional structure of a compatible system of solution functions, and is, in particular, loop-wisely solvable.

Moreover, the class of simple SCMs corresponds exactly with those SCMs that are loop-wisely uniquely solvable.

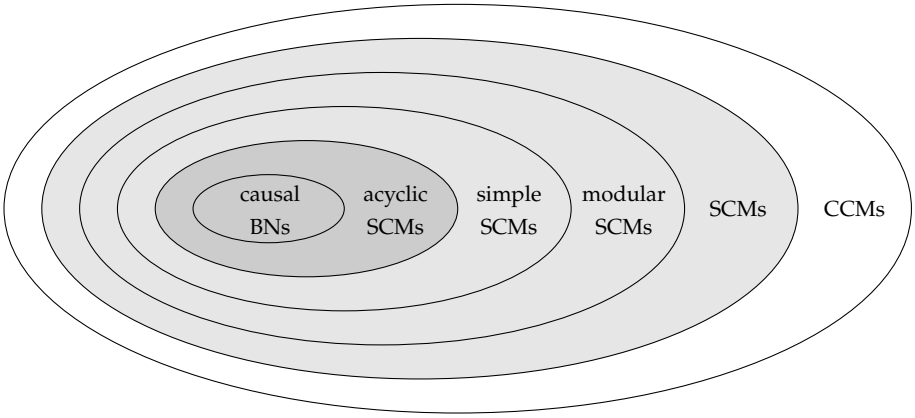


Figure 2.7: Overview of causal graphical models. The "gray" and "dark gray" areas contain all the causal graphical models that can be modeled by an SCM and an acyclic SCM, respectively.

Lemma 2.A.33. *An SCM $\mathcal{M}$ is simple if and only if it is loop-wisely uniquely solvable.*

In particular, for simple SCMs, or loop-wisely uniquely solvable SCMs, there always exists a compatible system of solution functions.

Proposition 2.A.34. *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be a simple SCM. Then every family of measurable solution functions $(\boldsymbol{g}_{\mathcal{O}})_{\mathcal{O} \in \mathcal{L}(\mathcal{G}(\mathcal{M}))}$, where $\boldsymbol{g}_{\mathcal{O}}$ is a measurable solution function of $\mathcal{M}$ w.r.t. $\mathcal{O}$, is a compatible system of solution functions.*

2.A.4 *Overview of causal graphical models*

Figure 2.7 gives an overview of the causal graphical models related to SCMs. The "gray" area contains all the causal graphical models that can be modeled by an SCM, by which we mean, that there exists an SCM that can describe all its observational and interventional distributions. The "dark gray" area contains all the causal graphical models which can be modeled by an acyclic SCM. Acyclic SCMs generalize causal Bayesian networks (causal BNs) (Pearl, 2009) to allow for latent confounders and to derive counterfactuals. Simple SCMs form a subclass of SCMs that extends acyclic SCMs to the cyclic setting, while preserving many of their convenient properties. Modular SCMs (Forré and Mooij, 2017) can be seen as SCMs that have an additional structure of compatible system of solution functions and contain, in particular, the class of simple SCMs. Forré and Mooij (2017) showed that modular SCMs satisfy various convenient properties, like marginalization and the general directed global Markov property. We show that for SCMs in general various of those properties still hold under certain solvability conditions. A generalization of SCMs, known as *causal constraints models (CCMs)*, has been proposed (Blom, Bongers, and Mooij, 2019) in order to completely model the causal semantics of the equilibrium solutions of a dynamical system given the initial conditions. This class of CCMs is rich enough to model the causal semantics of SCMs, but does not come

with a single graphical representation that provides both a Markov property and a causal interpretation (Blom, Diepen, and Mooij, 2021).

2.B (UNIQUE) SOLVABILITY PROPERTIES

In this appendix, we provide additional (unique) solvability properties for SCMs. In Appendix 2.B.1 we provide a sufficient condition of solvability w.r.t. (strict) subsets. In Appendix 2.B.2 we discuss how (unique) solvability is preserved under strict super- and subsets. In Appendix 2.B.3 we discuss how (unique) solvability is preserved under unions and intersections. The proofs of the theoretical results in this appendix are given in Appendix 2.E.

2.B.1 *Sufficient condition for solvability w.r.t. subsets*

For solvability w.r.t. a (strict) subset of $\mathcal{I}$ there exists a sufficient condition that is similar to the sufficient (and necessary) condition (2) in Theorem 2.3.2 in the sense that it is formulated in terms of the solutions of (a subset of) the structural equations, but no measurability is required.

Proposition 2.B.1 (Sufficient condition for solvability w.r.t. a subset)**.** *Let $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be an SCM and $\mathcal{O} \subseteq \mathcal{I}$ a subset. If for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\setminus\mathcal{O}}$ the topological space*

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

with the subspace topology induced by $\boldsymbol{\mathcal{X}}_{\mathcal{O}}$ is nonempty and σ-compact,[21] *then $\mathcal{M}$ is solvable w.r.t. $\mathcal{O}$.*

For many purposes, this condition of σ-compactness suffices since it contains for example all countable discrete spaces, every interval of the real line, and moreover all the Euclidean spaces. In particular, it suffices to prove a sufficient and necessary condition for unique solvability w.r.t. a subset, in terms of the solutions of a subset of the structural equations (see Theorem 2.3.6). For larger solution spaces, we refer the reader to (Kechris, 1995). For the class of linear SCMs (see Definition 2.C.1), we provide in Proposition 2.C.2 a sufficient and necessary condition for solvability w.r.t. a (strict) subset of $\mathcal{I}$.

2.B.2 *(Unique) solvability w.r.t. strict super- and subsets*

In general, (unique) solvability w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ does not imply (unique) solvability w.r.t. a strict superset $\mathcal{O} \subsetneq \mathcal{V} \subseteq \mathcal{I}$ nor w.r.t. a strict subset $\mathcal{W} \subsetneq \mathcal{O}$, as can be seen in the following example.

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

Example 2.B.2 (Solvability is not preserved under strict sub- or supersets)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *where the causal mechanism is given by*

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

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

However, in Proposition 2.3.10 we show that solvability w.r.t. $\mathcal{O}$ implies solvability w.r.t. every ancestral subset in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$.

2.B.3 *(Unique) solvability w.r.t. unions and intersections*

In general, (unique) solvability is not preserved under unions and intersections. The following example illustrates that (unique) solvability is in general not preserved under intersections.

Example 2.B.3 (Solvability is not preserved under intersections)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \emptyset, \mathbb{R}^3, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle$ *where the causal mechanism is given by*

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

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

Example 2.B.2 gives an example where (unique) solvability is not preserved under unions. Even, if we take the union of disjoint subsets, (unique) solvability is not preserved (see Example 2.2.4). Although, in general, unique solvability is not preserved under unions, we show next that unique solvability is preserved under the union of ancestral subsets, under the following assumptions.

Proposition 2.B.4 (Combining measurable solution functions on different sets)**.** *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be an SCM,* $\mathcal{O} \subseteq \mathcal{I}$ *a subset and* $\mathcal{A}, \tilde{\mathcal{A}} \subseteq \mathcal{O}$ *two ancestral subsets in* $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$*. If* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{A}$*,* $\tilde{\mathcal{A}}$ *and* $\mathcal{A} \cap \tilde{\mathcal{A}}$*, then* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{A} \cup \tilde{\mathcal{A}}$*.*

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

Corollary 2.B.5. *Let* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *be an SCM and* $\mathcal{O} \subseteq \mathcal{I}$ *a subset. Then* $\mathcal{M}$ *is ancestrally uniquely solvable w.r.t.* $\mathcal{O}$ *if and only if* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i)$ *for every* $i \in \mathcal{O}$*.*

2.C LINEAR SCMS

In this appendix, we provide some results about (unique) solvability and marginalization for linear SCMs. Linear SCMs form a special class of SCMs that has seen much attention in the literature (see, e.g., Bollen, 1989; Hyttinen, Eberhardt, and Hoyer, 2012). The proofs of the theoretical results in this appendix are given in Appendix 2.E.

Definition 2.C.1 (Linear SCM). *We call an SCM* $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \mathbb{R}^{\mathcal{I}}, \mathbb{R}^{\mathcal{J}}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^{\mathcal{J}}} \rangle$ *linear if each component of the causal mechanism is a linear combination of the endogenous and exogenous variables, that is*

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

where $i \in \mathcal{I}$, $B \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ *and* $\Gamma \in \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ *are matrices, and* $\mathbb{P}_{\mathbb{R}^{\mathcal{J}}}$ *is a product probability measure*[22] *on* $\mathbb{R}^{\mathcal{J}}$.

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

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

A nonzero coefficient B_{ij} for $i, j \in \mathcal{I}$ such that $i \neq j$ corresponds with a directed edge $j \to i$ in the (augmented) graph, and a coefficient $B_{ii} = 1$ for $i \in \mathcal{I}$ corresponds with a self-cycle $i \to i$ in the (augmented) graph of the SCM. A nonzero coefficient Γ_{ij} for $i \in \mathcal{I}$, $j \in \mathcal{J}$ with $\mathbb{P}_{\mathcal{E}_j}$ a nondegenerate probability distribution over $\mathbb{R}$ corresponds with a directed edge $j \to i$ in the augmented graph. A nonzero entry $(\Gamma\Gamma^T)_{ij}$ for $i, j \in \mathcal{I}$ with $i \neq j$ such that there exists a $k \in \mathcal{J}$ for which $\Gamma_{ik}, \Gamma_{jk} \neq 0$ and $\mathbb{P}_{\mathcal{E}_k}$ a nondegenerate probability distribution over $\mathbb{R}$ corresponds with a bidirected edge $i \leftrightarrow j$ in the graph of the SCM.

For linear SCMs, the solvability condition w.r.t. a subset, Definition 2.3.1, translates into a matrix condition. In order to state this condition we need to define the pseudoinverse (or the Moore-Penrose inverse) A^+ of a real matrix A (Penrose, 1955; Golub and Kahan, 1965). The *pseudoinverse of the matrix* A is defined by $A^+ := V\Sigma^+ U^*$, where $A = U\Sigma V^*$ is the singular value decomposition of A and Σ^+ is obtained by replacing each nonzero entry on the diagonal of Σ by its reciprocal (Golub and Kahan, 1965). One of its useful properties is that $AA^+A = A$.

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

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

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

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

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

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

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

Proposition 2.C.3 (Sufficient and necessary condition for unique solvability w.r.t. a subset for linear SCMs)**.** *Let $\mathcal{M}$ be a linear SCM, $\mathcal{L} \subseteq \mathcal{I}$ and $\mathcal{O} = \mathcal{I} \setminus \mathcal{L}$. Then $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if the matrix $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. Moreover, if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then the mapping $\boldsymbol{g}_{\mathcal{L}} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{L}}$ given by*

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

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

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

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

From Proposition 2.3.10 we know that an SCM is solvable w.r.t. $\mathcal{L}$ if and only if it is ancestrally solvable w.r.t. $\mathcal{L}$. In particular, this result also holds for linear SCMs. We saw in Example 2.3.11 that a similar result for unique solvability does not hold, that is, in general, it does not hold that unique solvability w.r.t. $\mathcal{L}$ implies ancestral unique solvability w.r.t. $\mathcal{L}$. For the class of linear SCMs we do have the following positive result.

Proposition 2.C.4 (Equivalent unique solvability conditions for linear SCMs)**.** *For a linear SCM $\mathcal{M}$ and a subset $\mathcal{L} \subseteq \mathcal{I}$ the following are equivalent:*

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

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

Proposition 2.C.5 (Marginalization of a linear SCM)**.** *Let $\mathcal{M}$ be a linear SCM and $\mathcal{L} \subseteq \mathcal{I}$ a subset of endogenous variables such that $\mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. Then there exists a marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ that is linear and with marginal causal mechanism $\tilde{\boldsymbol{f}} : \mathbb{R}^{\mathcal{O}} \times \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{O}}$ given by*

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

where $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$. Moreover, this marginalization respects the latent projection, that is,

$$\big(\mathcal{G}^a \circ \mathrm{marg}(\mathcal{L})\big)(\mathcal{M}) \subseteq \big(\mathrm{marg}(\mathcal{L}) \circ \mathcal{G}^a\big)(\mathcal{M})\,.$$

From Theorem 2.5.6 we know that $\mathcal{M}$ and its marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$ over $\mathcal{L}$ are observationally, interventionally and counterfactually equivalent w.r.t. $\mathcal{O}$. A similar result can also be found in (Hyttinen, Eberhardt, and Hoyer, 2012). In contrast to nonlinear SCMs, this class of linear SCMs has the convenient property that every marginalization of a model of this class respects the latent projection. Moreover, the subclass of simple linear SCMs is even closed under marginalization.

2.D EXAMPLES

In this appendix, we provide additional examples. In Appendix 2.D.1 we provide some examples of SCMs that describe the equilibrium states of certain feedback systems governed by (random) differential equations that motivated our study of cyclic SCMs (see Chapter 3 for further details). In Appendix 2.D.2 we provide additional examples that support the main text in Chapter 2.

2.D.1 *SCMs as equilibrium models*

In many systems occurring in the real world feedback loops between observed variables are present. For example, in economics, the price of a product may be a function of the demanded or supplied quantities, and vice versa; or in physics, two masses that are connected by a spring may exert forces on each other. Such systems are often described by a system of (random) differential equations. In Chapter 3 we show that SCMs are capable of modeling the causal semantics of the equilibrium states of such systems. For illustration purposes we provide the following toy example of interacting masses that are attached to springs.

Example 2.D.1 (Damped coupled harmonic oscillator)**.** *Consider a one-dimensional system of d point masses $m_i \in \mathbb{R}$ $(i = 1, \dots, d)$ with positions Q_i, which are coupled by springs, with spring constants $k_i > 0$ and equilibrium lengths $\ell_i > 0$ $(i = 0, \dots, d)$, under influence of friction with friction coefficients $b_i \in \mathbb{R}$ $(i = 1, \dots, d)$ and with fixed endpoints $Q_0 = 0$ and $Q_{d+1} = L > 0$ (see Figure 2.8 (top)). The equations of motion of this system are provided by the following differential equations*

$$\frac{d^2 Q_i}{dt^2} = \frac{k_i}{m_i}(Q_{i+1} - Q_i - \ell_i) + \frac{k_{i-1}}{m_i}(Q_{i-1} - Q_i + \ell_{i-1}) - \frac{b_i}{m_i}\frac{dQ_i}{dt} \qquad (i = 1, \dots, d)\,.$$

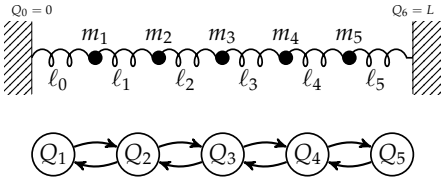


Figure 2.8: Damped coupled harmonic oscillator (top) and the graph of the SCM $\mathcal{M}$ that describes the positions of the masses at equilibrium (bottom) of Example 2.D.1 for $d = 5$.

The dynamics of the masses, in terms of the position, velocity and acceleration, is described by a single and separate equation of motion for each mass. Under friction, that is, $b_i > 0$ $(i = 1, \dots, d)$, there is a unique equilibrium position, where the sum of forces vanishes for each mass. If one starts out of equilibrium, for example, by moving one or several masses out of equilibrium, then the masses will start to oscillate and converge to their unique equilibrium position. At equilibrium (i.e., for $t \to \infty$) the velocity $\frac{dQ_i}{dt}$ and acceleration $\frac{d^2Q_i}{dt^2}$ of the masses vanish (i.e., $\frac{dQ_i}{dt}, \frac{d^2Q_i}{dt^2} \to 0$), and thus the following equation holds at equilibrium

$$0 = \frac{k_i}{m_i}(Q_{i+1} - Q_i - \ell_i) + \frac{k_{i-1}}{m_i}(Q_{i-1} - Q_i + \ell_{i-1}),$$

for each mass $(i = 1, \dots, d)$. Hence, for each mass $i = 1, \dots, d$ its equilibrium position Q_i is given by

$$Q_i = \frac{k_i(Q_{i+1} - \ell_i) + k_{i-1}(Q_{i-1} + \ell_{i-1})}{k_i + k_{i-1}}.$$

By considering the ℓ_i and k_i and L as fixed parameters, we arrive at a linear SCM (see Chapter 3 for more details about constructing an SCM from a dynamical system)

$$\mathcal{M} = \langle \{1, \dots, d\}, \emptyset, \mathbb{R}^d, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_{\mathbf{1}} \rangle,$$

where the causal mechanism $\boldsymbol{f}$ is given by

$$f_i(\boldsymbol{q}) = \frac{k_i(q_{i+1} - \ell_i) + k_{i-1}(q_{i-1} + \ell_{i-1})}{k_i + k_{i-1}}.$$

Alternatively, (some of) the parameters could be treated as exogenous variables instead. Its graph is depicted in Figure 2.8 (bottom). This SCM allows us to describe the equilibrium behavior of the system under perfect intervention. For example, when forcing the mass j to a fixed position $Q_j = \xi_j$ with $0 \leq \xi_j \leq L$, the equilibrium positions of the masses correspond to the solutions of the intervened model $\mathcal{M}_{\mathrm{do}(\{j\},\xi_j)}$. It is an easy exercise to show that $\mathcal{M}$ is a simple SCM by using Proposition 2.C.3.

Next, we show that the well known market equilibrium model from economics, which has been thoroughly discussed in the literature (see, e.g., Richardson and Robins, 2014), can be described by a (non-simple) SCM. This example illustrates how self-cycles enrich the class of SCMs.

Example 2.D.2 (Price, supply and demand)**.** *Let X_D denote the demand and X_S the supply of a quantity of a product. The price of the product is denoted by X_P. The following*

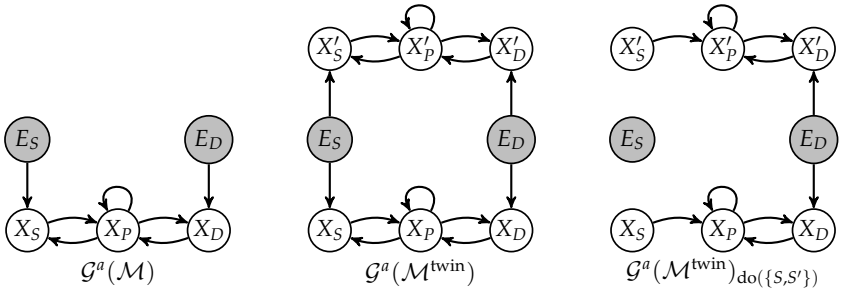


Figure 2.9: The augmented graph of the SCM $\mathcal{M}$ (left), its twin SCM $\mathcal{M}^{\text{twin}}$ (center) and the intervened twin SCM $(\mathcal{M}^{\text{twin}})_{\text{do}(\{S,S'\},(s,s'))}$ (right) of Examples 2.D.2 and 2.D.3.

system of differential equations describes how the demanded and supplied quantities are determined by the price, and how price adjustments occur in the market:

$$\begin{aligned} X_D &= \beta_D X_P + E_D \\ X_S &= \beta_S X_P + E_S \\ \frac{dX_P}{dt} &= X_D - X_S\,, \end{aligned}$$

where E_D and E_S are exogenous random influences on the demand and supply, respectively, $\beta_D < 0$ is the reciprocal of the slope of the demand curve, and $\beta_S > 0$ is the reciprocal of the slope of the supply curve. At the situation known as a "market equilibrium", the price is determined implicitly by the condition that demanded and supplied quantities should be equal, since $\frac{dX_P}{dt} = 0$ at equilibrium. Applying the results from Chapter 3 gives rise to a linear SCM $\mathcal{M} = \langle \{P, S, D\}, \{S, D\}, \mathbb{R}^3, \mathbb{R}^2, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\varepsilon}} \rangle$ at equilibrium with the causal mechanism defined by

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

Note how we use a self-cycle for P in order to implement the equilibrium equation $X_D = X_S$ as the causal mechanism for the price P.[23] *Moreover, $\mathcal{M}$ is uniquely solvable. Its augmented graph is depicted in Figure 2.9 (left).*

Next, we provide an example of how counterfactuals can be sensibly formulated for cyclic SCMs, namely for the price, supply and demand model at equilibrium.

Example 2.D.3 (Price, supply and demand at equilibrium)**.** *Consider the price, supply and demand model at equilibrium of Example 2.D.2 given by the SCM $\mathcal{M}$. As an example of a counterfactual query, consider*

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

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

which denotes the conditional distribution of X'_P given $X_P = p$ of a solution of the intervened twin model $\mathcal{M}^{\text{twin}}_{\text{do}(\{S,S'\},(s,s'))}$. In words: how would—ceteris paribus—*price have been distributed, had we intervened to set supplied quantities equal to s', given that actually we intervened to set supplied quantities equal to s and observed that this led to price p? A straightforward calculation shows that this counterfactual distribution of price is the Dirac measure on $x'_P = p + (s' - s)/\beta_D$. The augmented graphs of the SCM, its twin graph, and its intervened twin graph are depicted in Figure 2.9.*

2.D.2 *Additional examples*

In this subsection, we provide additional examples that support the main text in Chapter 2.

Section 2

Example 2.D.4 (Structural equations up to almost sure equality)**.** *Consider the SCM $\mathcal{M} = \langle \mathbf{1}, \mathbf{1}, \mathcal{X}, \mathcal{E}, f, \mathbb{P}_\mathcal{E} \rangle$ with $\mathcal{X} = \mathcal{E} = \{-1, 0, 1\}$, $\mathbb{P}_\mathcal{E}(\{-1\}) = \mathbb{P}_\mathcal{E}(\{1\}) = \frac{1}{2}$ and $f(x, e) = e^2 + e - 1$. Let $\tilde{\mathcal{M}}$ be the SCM $\mathcal{M}$ but with a different causal mechanism $\tilde{f}(x, e) = e$. Then the sets of solutions of the structural equations agree for both SCMs for $e \in \{-1, +1\}$, while they differ only for $e = 0$, which occurs with probability zero. Hence, a pair of random variables (X, E) is a solution of $\mathcal{M}$ if and only if it is a solution of $\tilde{\mathcal{M}}$.*

Example 2.D.5 (The for-all and for-almost-every quantifier do not commute in general)**.** *Consider the SCM $\mathcal{M} = \langle \mathbf{2}, \mathbf{1}, \boldsymbol{\mathcal{X}}, \mathcal{E}, \boldsymbol{f}, \mathbb{P}_\mathcal{E} \rangle$ with $\boldsymbol{\mathcal{X}} = (0,1)^2$, $\mathcal{E} = (0,1)$, the causal mechanism $\boldsymbol{f}$ given by*

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

and $\mathbb{P}_\mathcal{E} = \mathbb{P}^E$ with $E \sim \mathcal{U}(0,1)$. Define the property

$$P(\boldsymbol{x}, e) := \begin{cases} 1 & \text{if } \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, e) \text{ holds,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and for $\mathbb{P}_\mathcal{E}$-almost every $e \in \mathcal{E}$ the property $P(\boldsymbol{x}, e)$ holds, however for $\mathbb{P}_\mathcal{E}$-almost every $e \in \mathcal{E}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ the property $P(\boldsymbol{x}, e)$ does not hold, since for $\mathbb{P}_\mathcal{E}$-almost every $e \in \mathcal{E}$ the equation $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, e)$ does not hold for $x_1 = e$. Hence, in general, for a property $P(\boldsymbol{x}, e)$ we have that for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and for $\mathbb{P}_\mathcal{E}$-almost every $e \in \mathcal{E}$ $P(\boldsymbol{x}, e)$ does not imply for $\mathbb{P}_\mathcal{E}$-almost every $e \in \mathcal{E}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ $P(\boldsymbol{x}, e)$ (see Lemma 2.F.11 for additional properties of the for-almost-every quantifier).

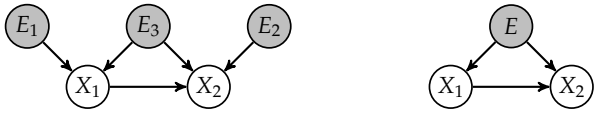


Figure 2.10: Augmented graphs of the SCMs $\mathcal{M}$ (left) and $\mathcal{M}^*$ (right) in Example 2.D.6. For SCM $\mathcal{M}^*$, the exogenous variable E consists of two real-valued components; the structural equation for X_1 depends only on the first, while the structural equation for X_2 depends only on the second component.

Example 2.D.6 (Representation of latent confounders)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{2}, \mathbf{3}, \mathbb{R}^2, \mathbb{R}^3, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}^3} \rangle$ *with causal mechanism given by*

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

and $\mathbb{P}_{\mathbb{R}^3}$ *the standard-normal distribution on* $\mathbb{R}^3$*; Figure 2.10 (left) shows the corresponding augmented graph. Then there exists no SCM* $\mathcal{M}^* = \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}^2, \boldsymbol{f}^*, \mathbb{P}^*_{\mathbb{R}^2} \rangle$ *that satisfies the following conditions:*

1. $\mathcal{M}^*$ *is interventionally equivalent to* $\mathcal{M}$*,*
2. *its structural equations have the form*

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

   where e_1^*, e_2^* *are the two components of* $e^* = (e_1^*, e_2^*) \in \mathbb{R}^2$*,*
3. *the function* $e_2^* \mapsto f_2^*(x_1, e_2^*)$ *is strictly monotonically increasing for all* $x_1 \in \mathbb{R}$*,*
4. *the cumulative distribution function* F_2^* *of the second component of* $\mathbb{P}^*_{\mathbb{R}^2}$ *is continuous and strictly monotonically increasing.*

The augmented graph of such an SCM is shown in Figure 2.10 (right).

The proof of this statement proceeds by contradiction. Assume that such an SCM $\mathcal{M}^*$ *exists. For any uniquely solvable SCM* $\bar{\mathcal{M}}$ *and any endogenous variable* i *appearing in* $\bar{\mathcal{M}}$*, we denote with* $F_{X_i}^{\bar{\mathcal{M}}}$ *the marginal cumulative distribution function of the* i^{th} *component of the observational distribution of* $\bar{\mathcal{M}}$*. For all* $\xi \in \mathbb{R}$*, we have for all* $x_2 \in \mathbb{R}$

$$F_{X_2}^{\mathcal{M}_{\mathrm{do}(\{1\},\xi)}}(x_2) = \mathbb{P}(\xi E_3 + E_2 \leq x_2) = \Phi\left(x_2 / \sqrt{1 + \xi^2}\right), \tag{2.1}$$

where Φ *denotes the (invertible) cdf of the standard-normal distribution. Now define* $\phi : \mathbb{R} \to \mathbb{R}$ *with* $\phi(e_2) := \Phi^{-1}(F_2^*(e_2))$ *and define the SCM* $\tilde{\mathcal{M}} := \langle \mathbf{2}, \mathbf{1}, \mathbb{R}^2, \mathbb{R}^2, \tilde{\boldsymbol{f}}, \tilde{\mathbb{P}}_{\mathbb{R}^2} \rangle$ *such that the causal mechanism* $\tilde{\boldsymbol{f}}$ *is given by*

$$\begin{aligned} \tilde{f}_1(e_1) &= f_1^*(e_1), \\ \tilde{f}_2(x_1, e_2) &= f_2^*(x_1, \phi^{-1}(e_2)), \end{aligned}$$

and $\tilde{\mathbb{P}}_{\mathbb{R}^2}$ is the push-forward measure of $\mathbb{P}^*_{\mathbb{R}^2}$ using $(\mathbb{I}_{\mathbb{R}}, \phi)$. Then, $\tilde{\mathcal{M}}$ is interventionally equivalent to $\mathcal{M}^*$ by construction, and the second component of $\tilde{\mathbb{P}}_{\mathbb{R}^2}$ has a standard-normal distribution. Let $(\tilde{X}_1, \tilde{X}_2, \tilde{E})$ be a solution of $\tilde{\mathcal{M}}$ and let us write $\tilde{E} = (\tilde{E}_1, \tilde{E}_2)$. Then, for all $\xi \in \mathbb{R}$ and $\tilde{e}_2 \in \mathbb{R}$,

$$F_{X_2}^{\tilde{\mathcal{M}}_{\mathrm{do}(\{1\},\xi)}}(\tilde{f}_2(\xi, \tilde{e}_2)) = \mathbb{P}(\tilde{f}_2(\xi, \tilde{E}_2) \leq \tilde{f}_2(\xi, \tilde{e}_2)) = \mathbb{P}(\tilde{E}_2 \leq \tilde{e}_2) = \Phi(\tilde{e}_2),$$

using that $\tilde{e}_2 \mapsto \tilde{f}_2(\xi, \tilde{e}_2)$, too, is strictly monotonically increasing for all ξ. This implies that, for all $\xi \in \mathbb{R}$ and $\tilde{e}_2 \in \mathbb{R}$,

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

where we used interventional equivalence of $\mathcal{M}$ and $\tilde{\mathcal{M}}$, and (2.1) for the second equality. Furthermore, $\tilde{X}_2 = \tilde{f}_2(\tilde{X}_1, \tilde{E}_2) = \sqrt{1 + \tilde{X}_1^2}\, \tilde{E}_2$ a.s., so $\tilde{E}_2 = \tilde{X}_2 / \sqrt{1 + \tilde{X}_1^2}$ a.s.. Now let $(X_1, X_2, E_1, E_2, E_3)$ be a solution of $\mathcal{M}$. By observational equivalence, $(\tilde{X}_1, \tilde{X}_2)$ has the same distribution as (X_1, X_2), and thus $\tilde{E}_2$ 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 $\tilde{E}_2$ has a standard-normal distribution as, for example, the mean of the right-hand side is nonzero.

Example 2.D.7 (Counterfactual density unidentifiable from observational and interventional densities (Dawid, 2002))**.** *Let $\rho \in \mathbb{R}$ and*

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

be the SCM with causal mechanism given by

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

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

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

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

Suppose in the actual world we did not assign treatment to a patient ($X_1 = 0$) and the outcome was $X_2 = c \in \mathbb{R}$. Consider the counterfactual query "What would the outcome have been, if we had assigned treatment to this patient?". We can answer this question by introducing a parallel counterfactual world that is modeled by the twin SCM $\mathcal{M}_\rho^{\mathrm{twin}}$, as

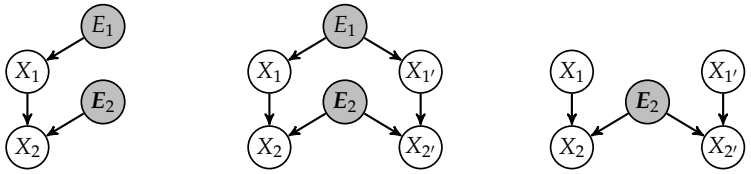


Figure 2.11: The augmented graph of the SCM $\mathcal{M}_\rho$ (left), its twin SCM $\mathcal{M}_\rho^{\text{twin}}$ (center) and the intervened twin SCM $(\mathcal{M}_\rho^{\text{twin}})_{\text{do}(\{1',1\},(1,0))}$ (right) of Example 2.D.7.

depicted in Figure 2.11. The counterfactual query then asks for $p(X_{2'} = x_{2'} \mid \text{do}(X_{1'} = 1, X_1 = 0), X_2 = c)$. *One can calculate that*

$$\begin{pmatrix} X_{2'} \\ X_2 \end{pmatrix} \mid \text{do}(X_{1'} = 1, X_1 = 0) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

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

Section 3

Example 2.D.8 (Mixtures of solutions are solutions)**.** *Let* $\mathcal{M} = \langle \mathbf{1}, \emptyset, \mathbb{R}, \mathbf{1}, f, \mathbb{P}_\mathbf{1} \rangle$ *be an SCM with causal mechanism* $f : \mathcal{X} \times \mathcal{E} \to \mathcal{X}$ *defined by* $f(x, e) = x - x^2 + 1$. *There exist only two measurable solution functions* $g_\pm : \mathcal{E} \to \mathcal{X}$ *for* $\mathcal{M}$, *defined by* $g_\pm(e) = \pm 1$. *Let* $X : \Omega \to \mathbb{R}$ *be a random variable that is a nontrivial mixture of point masses on* $\{-1, +1\}$. *Then* X *is a solution of* $\mathcal{M}$, *however neither* $g_+(E) = X$ *a.s., nor* $g_-(E) = X$ *a.s., for any random variable* E *such that* $\mathbb{P}^E = \mathbb{P}_\mathcal{E}$.

Example 2.D.9 (Solvability is not preserved under perfect intervention)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{2}, \emptyset, \mathbb{R}^2, \mathbf{1}, \boldsymbol{f}, \mathbb{P}_\mathbf{1} \rangle$ *with the following causal mechanism*

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

This SCM has a unique solution $(0, 1)$. *Doing a perfect intervention* $\text{do}(\{1\}, \xi_1)$ *for some* $\xi_1 \neq 0$, *however, leads to an intervened model* $\mathcal{M}_{\text{do}(\{1\},\xi_1)}$ *that is not solvable. Performing instead the perfect intervention* $\text{do}(\{2\}, \xi_2)$ *for some* $\xi_2 > 1$ *leads also to a nonuniquely solvable SCM* $\mathcal{M}_{\text{do}(\{2\},\xi_2)}$ *which has solutions with multiple induced distributions, for example,* $(X_1, X_2) = (\phi(\xi_2)\sqrt{\xi_2 - 1}, \xi_2)$ *with some measurable* $\phi : \mathbb{R} \to \{-1, +1\}$, *but also mixtures of those.*

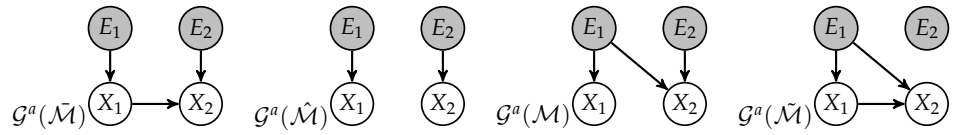


Figure 2.12: The augmented graphs of SCMs $\bar{\mathcal{M}}$, $\hat{\mathcal{M}}$, $\mathcal{M}$, and $\tilde{\mathcal{M}}$ that appear in Examples 2.4.4, 2.D.10, and 2.D.13.

Section 4

Example 2.D.10 (Counterfactually equivalent SCMs with different graphs)**.** *Consider the SCM* $\hat{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \{-1,1\}^2, \{-1,1\}^2, \hat{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *with causal mechanism given by* $\hat{f}_1(\boldsymbol{x}, \boldsymbol{e}) = e_1$ *and* $\hat{f}_2(\boldsymbol{x}, \boldsymbol{e}) = e_2$*, and* $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}^{\boldsymbol{E}}$ *with* $E_1, E_2 \sim \mathcal{U}(\{-1,1\})$ *uniformly distributed and* $E_1 \perp\!\!\!\perp E_2$*. Consider also the SCM* $\mathcal{M}$ *that is the same as* $\hat{\mathcal{M}}$ *except for its causal mechanism, which is given by* $f_1(\boldsymbol{x}, \boldsymbol{e}) = e_1$ *and* $f_2(\boldsymbol{x}, \boldsymbol{e}) = e_1 e_2$*. Then* $\mathcal{M}$ *and* $\hat{\mathcal{M}}$ *are counterfactually equivalent although* $\mathcal{G}(\mathcal{M})$ *is not equal to* $\mathcal{G}(\hat{\mathcal{M}})$ *(see Figure 2.12).*

Section 5

Example 2.D.11 (Marginalization condition of an SCM is not a necessary condition)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{4}, \mathbf{1}, \mathbb{R}^4, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ *with causal mechanism given by*

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

and $\mathbb{P}_{\mathbb{R}}$ *is the standard-normal measure on* $\mathbb{R}$*. This SCM is solvable w.r.t.* $\mathcal{L} = \{2, 4\}$*, but not uniquely solvable w.r.t.* $\mathcal{L}$*, and hence we cannot apply Definition 2.5.3 to* $\mathcal{L}$*. However, the SCM* $\tilde{\mathcal{M}}$ *on the endogenous variables* $\{1, 3\}$ *with the causal mechanism* $\tilde{\boldsymbol{f}}$ *given by* $\tilde{f}_1(\boldsymbol{x}, e) = e$ *and* $\tilde{f}_3(\boldsymbol{x}, e) = x_1$ *is counterfactually equivalent to* $\mathcal{M}$ *w.r.t.* $\{1, 3\}$*, which can be checked easily.*

Example 2.D.12 (Graph of the marginal SCM is a strict subgraph of the latent projection)**.** *Consider the SCM* $\mathcal{M} = \langle \mathbf{3}, \mathbf{1}, \mathbb{R}^3, \mathbb{R}, \boldsymbol{f}, \mathbb{P}_{\mathbb{R}} \rangle$ *with causal mechanism given by*

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

and take for $\mathbb{P}_{\mathbb{R}}$ *the standard-normal measure on* $\mathbb{R}$*. In contrast, to the (augmented) graph of* $\mathcal{M}$*, there is no directed path in the (augmented) graph of the marginal SCM* $\mathcal{M}_{\mathrm{marg}(\{3\})}$*.*

Section 7

Example 2.D.13 (Detecting a bidirected edge in the graph of an SCM)**.** *Consider the SCM* $\bar{\mathcal{M}} = \langle \mathbf{2}, \mathbf{2}, \{-1,1\}^2, \{-1,1\}^2, \bar{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ *with causal mechanism given by*

$$\begin{aligned} \bar{f}_1(\boldsymbol{x}, \boldsymbol{e}) &= e_1\,, \\ \bar{f}_2(\boldsymbol{x}, \boldsymbol{e}) &= x_1 e_2\,, \end{aligned}$$

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

$$\tilde{f}_1(\boldsymbol{x}, \boldsymbol{e}) = e_1\,,$$
$$\tilde{f}_2(\boldsymbol{x}, \boldsymbol{e}) = x_1 e_1\,.$$

See Figure 2.12 for their augmented graphs. For the SCM $\tilde{\mathcal{M}}$ *we observe that the marginal interventional distribution* $\mathbb{P}_{\tilde{\mathcal{M}}_{\mathrm{do}(\{1\},\xi_1)}}(X_2 = -1)$ *is not equal to the conditional distribution* $\mathbb{P}_{\tilde{\mathcal{M}}}(X_2 = -1 \mid X_1 = \xi_1)$ *for both* $\xi_1 = -1$ *and* $\xi_1 = 1$. *This observation suffices to identify the presence of the bidirected edge* $1 \leftrightarrow 2$ *in the graph* $\mathcal{G}(\tilde{\mathcal{M}})$. *For the SCM* $\bar{\mathcal{M}}$, *whose graph does not contain the bidirected edge* $1 \leftrightarrow 2$, *the marginal interventional distribution and conditional distribution coincide.*

2.E PROOFS

This appendix contains the proofs of all the theoretical results in the appendices 2.A, 2.B and 2.C, and the main text in Chapter 2. Some of the proofs will rely on the measure theoretic terminology and results of Appendix 2.F.

2.E.1 *Proofs of the appendices*

Appendix A

Proof of Lemma 2.A.5. It suffices to show that for every C-d-open walk between i and j in $\mathcal{G}$, there exists a C-d-open path between i and j in $\mathcal{G}$. Take a C-d-open walk $\pi = (i = i_0, \ldots, i_n = j)$. If a node ℓ occurs more than once in π, let i_j be the first occurrence of ℓ in π and i_k the last occurrence of ℓ in π. We now construct a new walk π' from π by removing the subwalk between i_j and i_k of π from π. It is easy to check that the new walk π' is still C-d-open. If ℓ is an endpoint on π', then i_j or i_k must be endpoint of π, and hence $\ell \notin C$. If ℓ is a non-endpoint non-collider on π', then also i_j or i_k must have been a non-endpoint non-collider on π, and hence $\ell \notin C$. If ℓ is a collider on π', then either (i) i_j or i_k are both colliders on π, and hence ℓ is ancestor of C in $\mathcal{G}$, or (ii) on the subwalk between i_j and i_k that was removed, there must be a directed path in $\mathcal{G}$ from i_j or i_k to a collider in $\mathrm{an}_{\mathcal{G}}(C)$, and hence, ℓ is in $\mathrm{an}_{\mathcal{G}}(C)$. The other nodes on π' cannot be responsible for C-d-blocking the walk, since they also occur (together with their adjacent edges) on π and they do not C-d-block π.

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

Proof of Theorem 2.A.7. The first case is a well known result. An elementary proof is obtained by noting that an acyclic system of structural equations trivially satisfies the local directed Markov property, and then apply (Lauritzen et al., 1990; Proposition 4), followed by applying the stability of d-separation with respect to (graphical) marginalization (Forré and Mooij, 2017; Lemma 2.2.15). Alternatively, the result

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

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

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

Proof of Proposition 2.A.12. This follows directly from the fact that the strongly connected components of $\mathcal{G}^a(\mathcal{M})$ form a DAG by Lemma 2.A.2 and that the directed edges in $\mathcal{G}^a(\mathrm{acy}(\mathcal{M}))$ by construction respect every topological ordering of that DAG. Both SCMs are observationally equivalent by construction. □

Proof of Proposition 2.A.14. This follows immediately from the Definitions 2.A.11 and 2.A.13. □

Proof of Lemma 2.A.17. It suffices to show that for every C-σ-open walk between i and j in $\mathcal{G}$, there exists a C-σ-open path between i and j in $\mathcal{G}$. Let $\pi = (i = i_0, \ldots, i_n = j)$ be a C-σ-open walk in $\mathcal{G}$. If a node ℓ occurs more than once in π, let i_j be the first node in π and i_k the last node in π that are in the same strongly connected component as ℓ. Since i_j and i_k are in the same strongly connected component, there are directed paths $i_j \to \cdots \to i_k$ and $i_k \to \cdots \to i_j$ in $\mathcal{G}$. We now construct a new walk π' from π by replacing the subwalk between i_j and i_k of π by a particular directed path between i_j and i_k: (i) If $k = n$, or if $k < n$ and $i_k \to i_{k+1}$ on π, we replace it by a shortest directed path $i_j \to \cdots \to i_k$, otherwise (ii) we replace it by a shortest directed path $i_j \leftarrow \cdots \leftarrow i_k$. We now show that the new walk π' is still C-σ-open.

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

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

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

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

Proof of Proposition 2.A.19. This follows directly as a special case of Corollary 2.8.4 in (Forré and Mooij, 2017). □

Proof of Theorem 2.A.21. An SCM $\mathcal{M}$ that is uniquely solvable w.r.t. each strongly connected component is uniquely solvable and hence, by Theorem 2.3.6, all its solutions have the same observational distribution. The last statement follows from the series of results Theorem 3.8.2, 3.8.11, Lemma 3.7.7 and Remark 3.7.2 in (Forré and Mooij, 2017). Alternatively, we give here a shorter proof: Under the stated conditions one can always construct the acyclification $\mathrm{acy}(\mathcal{M})$ which is observationally equivalent to $\mathcal{M}$ and is acyclic (see Proposition 2.A.12) and hence we can apply Theorem 2.A.7 to $\mathrm{acy}(\mathcal{M})$. Together with Proposition 2.A.14 and 2.A.19 this gives

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

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

Proof of Corollary 2.A.22. First observe that simplicity is preserved under both perfect intervention and the twin operation (see Proposition 2.8.2). Now the first statement follows from Theorem 2.A.21 if one takes into account the identities of Proposition 2.2.14 and 2.2.19. Similarly, the last statement follows from Theorem 2.A.7. □

Proof of Proposition 2.A.32. Let $\tilde{\mathcal{M}} =: \langle \mathcal{V}, \hat{\mathcal{H}}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ be the induced SCM. Observe that every loop $\mathcal{O} \in \mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))$ is a loop in $\mathcal{L}(\mathcal{G})$. Fix $\check{\boldsymbol{x}} \in \boldsymbol{\mathcal{X}}$ and $\check{\boldsymbol{e}} \in \boldsymbol{\mathcal{E}}$. For every $\mathcal{O} \in \mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))$, define

$$I_{\mathcal{O}} := (\mathrm{pa}_{\mathcal{G}}(\mathcal{O}) \setminus \mathcal{O}) \setminus (\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}) \subseteq \tilde{\mathcal{I}}$$

and

$$J_{\mathcal{O}} := \{\mathcal{F} \in \tilde{\mathcal{J}} \,:\, \mathcal{F} \cap \mathcal{O} \neq \emptyset\} \setminus \mathrm{pa}(\mathcal{O}) \subseteq \tilde{\mathcal{J}}\,.$$

Now, define the family of measurable mappings $(\tilde{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))}$, where the mapping $\tilde{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ is given by

$$\tilde{g}_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) := g_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \check{x}_{I_{\mathcal{O}}}, e_{\mathrm{pa}(\mathcal{O})}, \check{e}_{J_{\mathcal{O}}})$$

where $x_{\mathrm{pa}_{\mathcal{G}}(\mathcal{O})\setminus\mathcal{O}} = (x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, \check{x}_{I_{\mathcal{O}}})$ and $\widehat{e}_{\mathcal{O}} = (e_{\mathrm{pa}(\mathcal{O})}, \check{e}_{J_{\mathcal{O}}})$. Observe that from the definition of the parents (see Definition 2.2.6) it follows that for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$ we have

$$x_{\mathcal{O}} = \tilde{f}_{\mathcal{O}}(x_{\setminus I_{\mathcal{O}}}, \check{x}_{I_{\mathcal{O}}}, e_{\setminus J_{\mathcal{O}}}, \check{e}_{J_{\mathcal{O}}}) \quad \iff \quad x_{\mathcal{O}} = \tilde{f}_{\mathcal{O}}(x, e)\,.$$

This, together with the fact that the family of mappings $(g_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G})}$ is a compatible system of solution functions, implies that for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$ we have

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

Hence, $\iota(\widehat{\mathcal{M}})$ is loop-wisely solvable and thus $(\tilde{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))}$ is a family of measurable solution functions. In particular, for all $\mathcal{O}, \tilde{\mathcal{O}} \in \mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))$ with $\tilde{\mathcal{O}} \subseteq \mathcal{O}$ and for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$ we have

$$x_{\mathcal{O}} = \tilde{g}_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \quad \implies \quad x_{\tilde{\mathcal{O}}} = \tilde{g}_{\tilde{\mathcal{O}}}(x_{\mathrm{pa}(\tilde{\mathcal{O}})\setminus\tilde{\mathcal{O}}}, e_{\mathrm{pa}(\tilde{\mathcal{O}})})\,.$$

From this we conclude that $(\tilde{g}_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G}(\tilde{\mathcal{M}}))}$ is a compatible system of solution functions. □

Proof of Lemma 2.A.33. Suppose $\mathcal{M}$ is loop-wisely uniquely solvable and consider a subset $\mathcal{O} \subseteq \mathcal{I}$. Consider the induced subgraph $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ of $\mathcal{G}^a(\mathcal{M})$ on the nodes $\mathcal{O}$. Then every strongly connected component of $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ is an element of $\mathcal{L}(\mathcal{G}(\mathcal{M}))$. Let $\mathcal{C}$ be such a strongly connected component in $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$, and let $g_{\mathcal{C}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{C})\setminus\mathcal{C}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{C})} \to \boldsymbol{\mathcal{X}}_{\mathcal{C}}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{C}$. Since $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ partitions into strongly connected components, we can recursively (by following a topological ordering of the DAG $\mathcal{G}^a(\mathcal{M})^{\mathrm{sc}}_{\mathcal{O}}$ from Lemma 2.A.2) insert these mappings into each other to obtain a mapping $g_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ that makes $\mathcal{M}$ uniquely solvable w.r.t. $\mathcal{O}$. □

Proof of Proposition 2.A.34. Let $(g_{\mathcal{O}})_{\mathcal{O}\in\mathcal{L}(\mathcal{G}(\mathcal{M}))}$ be any family of measurable solution functions, where $g_{\mathcal{O}}$ is measurable solution function of $\mathcal{M}$ w.r.t. $\mathcal{O}$. Then, for $\mathcal{O}, \tilde{\mathcal{O}} \in \mathcal{L}(\mathcal{G}(\mathcal{M}))$ such that $\tilde{\mathcal{O}} \subseteq \mathcal{O}$, we have that for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$

$$x_{\mathcal{O}} = f_{\mathcal{O}}(x, e) \quad \implies \quad x_{\tilde{\mathcal{O}}} = f_{\tilde{\mathcal{O}}}(x, e)\,.$$

This implies that for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$

$$x_{\mathcal{O}} = g_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \quad \implies \quad x_{\tilde{\mathcal{O}}} = g_{\tilde{\mathcal{O}}}(x_{\mathrm{pa}(\tilde{\mathcal{O}})\setminus\tilde{\mathcal{O}}}, e_{\mathrm{pa}(\tilde{\mathcal{O}})})\,.$$

□

Proof of Corollary 2.8.5. This follows directly from Proposition 2.7.1 and 2.7.2. □

Appendix B

Proof of Proposition 2.B.1. Let $\tilde{f} : \boldsymbol{\mathcal{E}} \times \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be the causal mechanism of a structurally minimal SCM that is equivalent to $\mathcal{M}$ (see Proposition 2.2.11). In particular, for any $\boldsymbol{\epsilon}_{\setminus \mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{E}}_{\setminus \mathrm{pa}(\mathcal{O})}$ and $\boldsymbol{\xi}_{\setminus \mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{X}}_{\setminus \mathrm{pa}(\mathcal{O})}$, we have that for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, $\tilde{f}(\boldsymbol{x}, \boldsymbol{e}) = \tilde{f}(\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{\xi}_{\setminus \mathrm{pa}(\mathcal{O})}, \boldsymbol{e}_{\mathrm{pa}(\mathcal{O})}, \boldsymbol{\epsilon}_{\setminus \mathrm{pa}(\mathcal{O})})$. This means that we may also consider $\tilde{f}$ as a mapping $\tilde{f} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}$.

Consider the set

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

By similar reasoning as in the proof of Theorem 2.3.2, $\tilde{\boldsymbol{\mathcal{S}}}$ is measurable.

By assumption, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\setminus\mathcal{O}}$ the space $\{\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}} : \boldsymbol{x}_{\mathcal{O}} = f_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e})\}$ is nonempty and σ-compact. By applying Lemma 2.F.10 to the canonical projection $pr_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}} : \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}$ and using the equivalence of f and $\tilde{f}$, we obtain that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}}$-almost every $\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}$ and for all $\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}$ the space

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

is nonempty and σ-compact.

The second measurable selection theorem, Theorem 2.F.9, now implies that there exists a measurable $g_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}}$-almost every $\boldsymbol{e}_{\mathrm{pa}(\mathcal{O})} \in \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})}$ and for all $\boldsymbol{x}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}$

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

Once more applying Lemma 2.F.10, we obtain that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

Proof of Proposition 2.B.4. Without loss of generality, we assume that $\mathcal{M}$ is structurally minimal (see Proposition 2.2.11). Define $\mathcal{C} := \mathcal{A} \cap \tilde{\mathcal{A}}$ and $\mathcal{D} := \mathcal{A} \cup \tilde{\mathcal{A}}$. Let $g_{\mathcal{A}}$, $g_{\tilde{\mathcal{A}}}$ be measurable solution functions for $\mathcal{M}$ w.r.t. $\mathcal{A}$ and $\tilde{\mathcal{A}}$, respectively. Note that $\mathrm{pa}(\mathcal{C}) \setminus \mathcal{C} \subseteq \mathrm{pa}(\mathcal{A}) \setminus \mathcal{A}$ and similarly $\mathrm{pa}(\mathcal{C}) \setminus \mathcal{C} \subseteq \mathrm{pa}(\tilde{\mathcal{A}}) \setminus \tilde{\mathcal{A}}$. Indeed, for $c \in \mathrm{pa}(\mathcal{C})$: if $c \in \mathcal{O}$ then $c \in \mathcal{C}$ because $\mathcal{A}$ and $\tilde{\mathcal{A}}$ are both ancestral in $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$, while if $c \notin \mathcal{O}$ then $c \notin \mathcal{A}$ and $c \notin \tilde{\mathcal{A}}$. Hence by Lemma 2.E.1, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

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

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

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

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

Proof of Corollary 2.B.5. It suffices to show the implication to the left. We have to show that $\mathcal{M}$ is uniquely solvable w.r.t. each ancestral subset of $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$. The proof proceeds via induction with respect to the size of the ancestral subset. For ancestral subsets of size 0, the claim is trivially true. Ancestral subsets of size 1 must be of the form $\{i\} = \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i)$ for $i \in \mathcal{O}$ and hence the claim is true by assumption. Assume that the claim holds for all ancestral subsets of size $\leq n$. Let $\mathcal{A}$ be an ancestral subset of $\mathcal{G}(\mathcal{M})_{\mathcal{O}}$ of size $n+1$. If $\mathcal{A} = \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i)$ for some $i \in \mathcal{O}$ then the claim holds for $\mathcal{A}$ by assumption. Otherwise, $\mathcal{A} = \bigcup_{i\in\mathcal{A}} \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i)$ is a union of ancestral subsets of size $\leq n$. Choose distinct elements $\{i_1, \ldots, i_k\} \subseteq \mathcal{A}$ where k is the smallest integer such that $\bigcup_{j=1}^{k} \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i_j) = \mathcal{A}$. By applying Proposition 2.B.4 to $\bigcup_{j=1}^{k-1} \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i_j)$ and $\mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(i_k)$, thereby noting that the intersection of these two sets is an ancestral subset of size $\leq n$ and making use of the induction hypothesis, we arrive at the conclusion that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{A}$. □

Appendix C

Proof of Proposition 2.C.2. Let $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$. For $\boldsymbol{x}_{\mathcal{L}} \in \boldsymbol{\mathcal{X}}$,

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

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

Proof of Proposition 2.C.3. $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ the linear system of equations

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

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

Proof of Proposition 2.C.4. It suffices to show (1) $\implies$ (2) and (1) $\iff$ (3). We start by showing that (1) $\implies$ (2). Let $\mathcal{V} \subseteq \mathcal{L}$ and denote $\mathcal{U} := \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{L}}}(\mathcal{V})$, then we need to show that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{U}$. From Proposition 2.C.3 we know that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if the matrix $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. The matrix $A_{\mathcal{L}\mathcal{L}}$ is invertible if and only if the rows of $A_{\mathcal{L}\mathcal{L}}$ are all linearly independent. In particular, the rows of $A_{\mathcal{U}\mathcal{L}}$ are all linearly independent. Because $A_{\mathcal{U}\mathcal{L}} = [A_{\mathcal{U}\mathcal{U}}\, Z_{\mathcal{U}\mathcal{L}}]$, where $Z_{\mathcal{U}\mathcal{L}}$ is the zero matrix, we know that the rows of $A_{\mathcal{U}\mathcal{U}} = \mathbb{I}_{\mathcal{U}} - B_{\mathcal{U}\mathcal{U}}$ are also all linearly independent, and hence $A_{\mathcal{U}\mathcal{U}}$ is invertible.

Next, we show that (1) $\iff$ (3). Observe that the strongly connected components of $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$ form a partition of the set $\mathcal{L}$ and that the directed mixed graph $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$ and the directed graph $\mathcal{G}^{a}(\mathcal{M})_{\mathcal{L}}$ have the same strongly connected components. Because, by Lemma 2.A.2, the graph of strongly connected components $\mathcal{G}^{\mathrm{sc}}$ of the directed graph $\mathcal{G}^{a}(\mathcal{M})_{\mathcal{L}}$ is a DAG, the square matrix $B_{\mathcal{L}\mathcal{L}}$ can be permuted to an upper triangular block matrix $\tilde{B}_{\mathcal{L}\mathcal{L}}$, where for each diagonal block $\tilde{B}_{\mathcal{V}\mathcal{V}}$ of $\tilde{B}_{\mathcal{L}\mathcal{L}}$ the set of nodes $\mathcal{V}$ is a strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$.

Without loss of generality we assume now that $B_{\mathcal{L}\mathcal{L}}$ is an upper triangular block matrix. From Proposition 2.C.3 it follows that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$ if and only if the matrix $A_{\mathcal{L}\mathcal{L}} = \mathbb{I}_{\mathcal{L}} - B_{\mathcal{L}\mathcal{L}}$ is invertible. Because $B_{\mathcal{L}\mathcal{L}}$ is an upper triangular block matrix, we know that $A_{\mathcal{L}\mathcal{L}}$ is an upper triangular block matrix, where for each diagonal block $A_{\mathcal{V}\mathcal{V}}$ of $A_{\mathcal{L}\mathcal{L}}$ the set of nodes $\mathcal{V}$ is a strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$. Since an upper triangular block matrix $A_{\mathcal{L}\mathcal{L}}$ is invertible if and only if every diagonal block in $A_{\mathcal{L}\mathcal{L}}$ is invertible, we have that $\mathcal{M}$ is uniquely solvable w.r.t.

$\mathcal{L}$ if and only if $\mathcal{M}$ is uniquely solvable w.r.t. each strongly connected component in $\mathcal{G}(\mathcal{M})_{\mathcal{L}}$. □

Proof of Proposition 2.C.5. By the definition of marginalization and Proposition 2.C.3 the marginal causal mechanism $\tilde{\boldsymbol{f}}$ is given by

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

From Propositions 2.C.4 and 2.5.11 it follows that the marginalization respects the latent projection. □

2.E.2 *Proofs of the main text*

Section 2

Proof of Proposition 2.2.11. Let $i \in \mathcal{I}$. Note that Definition 2.2.6 can alternatively be formulated as follows: for $k \in \mathcal{I} \cup \mathcal{J}$, $k \notin \mathrm{pa}(i)$ if and only if there exists a measurable mapping $\hat{f}_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

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

and either $k \in \mathcal{I}$ and there exists $\hat{x}_k \in \mathcal{X}_k$ such that $\hat{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \hat{f}_i(\boldsymbol{x}_{\setminus k}, \hat{x}_k, \boldsymbol{e})$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}}$, or $k \in \mathcal{J}$ and there exists $\hat{e}_k \in \mathcal{E}_k$ such that $\hat{f}_i(\boldsymbol{x}, \boldsymbol{e}) = \hat{f}_i(\boldsymbol{x}, \boldsymbol{e}_{\setminus k}, \hat{e}_k)$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{e} \in \boldsymbol{\mathcal{E}}$. By repeatedly applying (this formulation of) Definition 2.2.6 to all $k \notin \mathrm{pa}(i)$, we obtain the existence of a measurable mapping $\tilde{f}_i : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \mathcal{X}_i$ and $\hat{\boldsymbol{x}}_{\setminus \mathrm{pa}(i)} \in \boldsymbol{\mathcal{X}}_{\setminus \mathrm{pa}(i)}$, $\hat{\boldsymbol{e}}_{\setminus \mathrm{pa}(i)} \in \boldsymbol{\mathcal{E}}_{\setminus \mathrm{pa}(i)}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$,

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

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

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

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

Proof of Proposition 2.2.14. The $\mathrm{do}(I, \boldsymbol{\xi}_I)$ operation on $\mathcal{M}$ completely removes the functional dependence on $\boldsymbol{x}$ and $\boldsymbol{e}$ from the f_i components for $i \in I$ and hence the corresponding incoming directed and bidirected edges on nodes in I from the (augmented) graph. □

Proof of Proposition 2.2.15. The first statement follows from Definitions 2.2.12 and 2.2.13. For the second statement, note that a perfect intervention can only remove parental relations, and therefore will never introduce a cycle. □

Proof of Proposition 2.2.19. This follows directly from Definitions 2.2.17 and 2.2.18. □

Proof of Proposition 2.2.20. The additional edges introduced by the twin operation cannot lead to a directed cycle involving both copied and original nodes, because there are no edges pointing from copied nodes to original nodes (i.e., of the form $i' \to v$ with $i' \in I'$ and $v \in \mathcal{V}$). Directed cycles involving only original nodes are absent by assumption, and directed cycles involving only copied nodes as well since they would correspond with a directed cycle in the original directed graph. □

Proof of Proposition 2.2.21. It suffices to prove the property for directed graphs, since the property for SCMs follows directly from Definitions 2.2.12 and 2.2.17.

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

Section 3

Proof of Theorem 2.3.2. First we define the solution space $\boldsymbol{\mathcal{S}}(\mathcal{M})$ of $\mathcal{M}$ by

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

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

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

is an analytic set because the projection $\boldsymbol{pr}_{\boldsymbol{\mathcal{E}}} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{E}}$ is a measurable mapping between standard measurable spaces (Lemma 2.F.3).

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

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

By Lemma 2.F.6, $\boldsymbol{\mathcal{A}}$ is $\mathbb{P}^{\boldsymbol{E}}$-measurable because it is analytic, and we can write $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{B}} \cup \boldsymbol{\mathcal{N}}$ with $\boldsymbol{\mathcal{B}} \subseteq \boldsymbol{\mathcal{E}}$ measurable and $\boldsymbol{\mathcal{N}}$ a $\mathbb{P}^{\boldsymbol{E}}$-null set. Hence $\boldsymbol{E}^{-1}(\boldsymbol{\mathcal{A}}) = \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{B}}) \cup \boldsymbol{E}^{-1}(\boldsymbol{\mathcal{N}})$ where $\boldsymbol{E}^{-1}(\boldsymbol{\mathcal{N}})$ is a $\mathbb{P}$-null set. Therefore,

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

which implies that $\mathbb{P}(\boldsymbol{E}^{-1}(\mathcal{B})) = 1$. Hence, $\mathcal{E} \setminus \mathcal{A}$ is a $\mathbb{P}_{\mathcal{E}}$-null set. In other words, for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ the structural equations $x = f(x, e)$ have a solution $x \in \mathcal{X}$, that is, (2) holds.

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

$$x = g(e) \quad \implies \quad x = f(x, e)\,,$$

which we call property (A). Let $\tilde{f} : \mathcal{E} \times \mathcal{X} \to \mathcal{X}$ be the causal mechanism of a structurally minimal SCM that is equivalent to $\mathcal{M}$ (see Proposition 2.2.11). In particular, for any $e_{\setminus \mathrm{pa}(\mathcal{I})} \in \mathcal{E}_{\setminus \mathrm{pa}(\mathcal{I})}$, we have that $\tilde{f}(x, e) = \tilde{f}(x, e_{\mathrm{pa}(\mathcal{I})}, e_{\setminus \mathrm{pa}(\mathcal{I})})$ for all $x \in \mathcal{X}$ and all $e \in \mathcal{E}$. This means that we may also consider $\tilde{f}$ as a mapping $\tilde{f} : \mathcal{X} \times \mathcal{E}_{\mathrm{pa}(\mathcal{I})} \to \mathcal{X}$. By applying Lemma 2.F.10 to the canonical projection $pr_{\mathcal{E}_{\mathrm{pa}(\mathcal{I})}} : \mathcal{E} \to \mathcal{E}_{\mathrm{pa}(\mathcal{I})}$ and using the equivalence of f and $\tilde{f}$, we obtain that for $\mathbb{P}_{\mathcal{E}_{\mathrm{pa}(\mathcal{I})}}$-almost all $e_{\mathrm{pa}(\mathcal{I})} \in \mathcal{E}_{\mathrm{pa}(\mathcal{I})}$ there exists $x \in \mathcal{X}$ with $x = \tilde{f}(x, e_{\mathrm{pa}(\mathcal{I})})$. By applying the implication (2) $\implies$ (A) to $\mathcal{E}_{\mathrm{pa}(\mathcal{I})}$ and $\tilde{f}$, we conclude the existence of a measurable $g : \mathcal{E}_{\mathrm{pa}(\mathcal{I})} \to \mathcal{X}$ such that for $\mathbb{P}_{\mathcal{E}_{\mathrm{pa}(\mathcal{I})}}$-almost all $e_{\mathrm{pa}(\mathcal{I})} \in \mathcal{E}_{\mathrm{pa}(\mathcal{I})}$, $g(e_{\mathrm{pa}(\mathcal{I})}) = \tilde{f}(g(e_{\mathrm{pa}(\mathcal{I})}), e_{\mathrm{pa}(\mathcal{I})})$. Once more using Lemma 2.F.10, we obtain that for $\mathbb{P}_{\mathcal{E}}$-almost all $e \in \mathcal{E}$, $g(e_{\mathrm{pa}(\mathcal{I})}) = f(g(e_{\mathrm{pa}(\mathcal{I})}), e)$. In other words, (3) holds.

Lastly, suppose that (3) holds, that is there exists a measurable solution function $g : \mathcal{E}_{\mathrm{pa}(\mathcal{I})} \to \mathcal{X}$. Then the measurable mappings $\boldsymbol{E} : \mathcal{E} \to \mathcal{E}$ and $\boldsymbol{X} : \mathcal{E} \to \mathcal{X}$, defined by $\boldsymbol{E}(e) := e$ and $\boldsymbol{X}(e) := g(e_{\mathrm{pa}(\mathcal{I})})$, respectively, define a pair of random variables $(\boldsymbol{X}, \boldsymbol{E})$ such that $\boldsymbol{X} = f(\boldsymbol{X}, \boldsymbol{E})$ holds a.s. and hence $(\boldsymbol{X}, \boldsymbol{E})$ is a solution. Hence (1) holds. □

Proof of Proposition 2.3.4. Let $\tilde{f} : \mathcal{E} \times \mathcal{X} \to \mathcal{X}$ be the causal mechanism of a structurally minimal SCM $\tilde{\mathcal{M}}$ that is equivalent to $\mathcal{M}$ (see Proposition 2.2.11). For a subset $\mathcal{O} \subseteq \mathcal{I}$ consider the induced subgraph $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ of the augmented graph $\mathcal{G}^a(\mathcal{M})$ on $\mathcal{O}$. Then the acyclicity of $\mathcal{G}^a(\mathcal{M})$ implies that the induced subgraph $\mathcal{G}^a(\mathcal{M})_{\mathcal{O}}$ is acyclic, and hence there exists a topological ordering on the nodes $\mathcal{O}$. We can substitute the components $\tilde{f}_i$ of the causal mechanism $\tilde{f}$ for $i \in \mathcal{O}$ into each other along this topological ordering. This gives a measurable solution function $g_{\mathcal{O}} : \mathcal{X}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \mathcal{E}_{\mathrm{pa}(\mathcal{O})} \to \mathcal{X}_{\mathcal{O}}$ for $\tilde{\mathcal{M}}$, and hence for $\mathcal{M}$. It is clear from the acyclic structure that this mapping $g_{\mathcal{O}}$ is independent of the choice of the topological ordering and is the only solution function for $\mathcal{M}$. Therefore, $\tilde{\mathcal{M}}$ is uniquely solvable w.r.t. $\mathcal{O}$, and so is $\mathcal{M}$. □

Proof of Proposition 2.3.7. This follows immediately from Definitions 2.2.7 and 2.3.3. □

Proof of Theorem 2.3.6. Suppose that (1) holds. By Proposition 2.B.1 there exists a measurable solution function $g_{\mathcal{O}} : \mathcal{X}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \mathcal{E}_{\mathrm{pa}(\mathcal{O})} \to \mathcal{X}_{\mathcal{O}}$ for $\mathcal{M}$ w.r.t. $\mathcal{O}$. Then for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x_{\setminus\mathcal{O}} \in \mathcal{X}_{\setminus\mathcal{O}}$ we have that $g_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})})$ is a solution of $x_{\mathcal{O}} = f_{\mathcal{O}}(x, e)$. Hence, because of (1), for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and

for all $x_{\setminus\mathcal{O}} \in \mathcal{X}_{\setminus\mathcal{O}}$ we have that $x_{\mathcal{O}} = f_{\mathcal{O}}(x, e)$ implies $x_{\mathcal{O}} = g_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})})$. Thus, $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{O}$, that is, (2) holds.

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

$$x_{\mathcal{O}} = g_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \quad \iff \quad x_{\mathcal{O}} = f_{\mathcal{O}}(x, e)\,.$$

This implies (1).

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

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

The existence of a solution for $\mathcal{M}$ follows directly from Theorem 2.3.2. Each solution $(X, E) : \Omega \to \mathcal{X} \times \mathcal{E}$ of $\mathcal{M}$ satisfies $X(\omega) = f(X(\omega), E(\omega))$ $\mathbb{P}$-a.s.. In addition, it satisfies $E(\omega) \in B$ $\mathbb{P}$-a.s., since $\mathbb{P} \circ E^{-1} = \mathbb{P}_{\mathcal{E}}$. Hence, it satisfies $X(\omega) = g(E(\omega)_{\mathrm{pa}(\mathcal{I})})$ $\mathbb{P}$-a.s.. Thus for every solution (X, E) the associated observational distribution is the push-forward of $\mathbb{P}_{\mathcal{E}}$ under $g \circ pr_{\mathrm{pa}(\mathcal{I})}$. □

Proof of Proposition 2.3.8. Let $g_{\mathcal{O}} : \mathcal{X}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \mathcal{E}_{\mathrm{pa}(\mathcal{O})} \to \mathcal{X}_{\mathcal{O}}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{O}$. Then the mapping $\tilde{g}_{\mathcal{O}\cup I} : \mathcal{E}_{\mathrm{pa}(\mathcal{O})} \to \mathcal{X}_{\mathcal{O}\cup I}$ defined by $\tilde{g}_{\mathcal{O}\cup I}(e_{\mathrm{pa}(\mathcal{O})}) := (g_{\mathcal{O}}(\xi_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}), \xi_I)$ is a measurable solution function for the SCM $\mathcal{M}_{\mathrm{do}(I,\xi_I)}$ w.r.t. $\mathcal{O} \cup I$. If $\mathcal{M}$ is (uniquely) solvable w.r.t. $\mathcal{O}$, then it follows that $\mathcal{M}_{\mathrm{do}(I,\xi_I)}$ is (uniquely) solvable w.r.t. $\mathcal{O} \cup I$. □

Proof of Proposition 2.3.10. It suffices to show that solvability of $\mathcal{M}$ w.r.t. $\mathcal{O}$ implies ancestral solvability w.r.t. $\mathcal{O}$. Solvability of $\mathcal{M}$ w.r.t. $\mathcal{O}$ implies that there exists a measurable mapping $g_{\mathcal{O}} : \mathcal{X}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \mathcal{E}_{\mathrm{pa}(\mathcal{O})} \to \mathcal{X}_{\mathcal{O}}$ such that for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$

$$x_{\mathcal{O}} = g_{\mathcal{O}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \quad \implies \quad x_{\mathcal{O}} = f_{\mathcal{O}}(x, e)\,.$$

Let $\tilde{f} : \mathcal{E} \times \mathcal{X} \to \mathcal{X}$ be the causal mechanism of a structurally minimal SCM $\tilde{\mathcal{M}}$ that is equivalent to $\mathcal{M}$ (see Proposition 2.2.11). Let $\mathcal{P} := \mathrm{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{O}}}(\mathcal{A})$ for some $\mathcal{A} \subseteq \mathcal{O}$. Then for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$ and for all $x \in \mathcal{X}$

$$\begin{cases} x_{\mathcal{P}} &= (g_{\mathcal{O}})_{\mathcal{P}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \\ x_{\mathcal{O}\setminus\mathcal{P}} &= (g_{\mathcal{O}})_{\mathcal{O}\setminus\mathcal{P}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \end{cases} \implies \begin{cases} x_{\mathcal{P}} &= \tilde{f}_{\mathcal{P}}(x_{\mathrm{pa}(\mathcal{P})}, e_{\mathrm{pa}(\mathcal{P})}) \\ x_{\mathcal{O}\setminus\mathcal{P}} &= \tilde{f}_{\mathcal{O}\setminus\mathcal{P}}(x_{\mathrm{pa}(\mathcal{O}\setminus\mathcal{P})}, e_{\mathrm{pa}(\mathcal{O}\setminus\mathcal{P})})\,. \end{cases}$$

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

$$x_{\mathcal{P}} = (g_{\mathcal{O}})_{\mathcal{P}}(x_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}}, e_{\mathrm{pa}(\mathcal{O})}) \quad \implies \quad x_{\mathcal{P}} = \tilde{f}_{\mathcal{P}}(x_{\mathrm{pa}(\mathcal{P})}, e_{\mathrm{pa}(\mathcal{P})})\,.$$

This implies that the mapping $(g_{\mathcal{O}})_{\mathcal{P}}$ cannot depend on elements different from $\mathrm{pa}(\mathcal{P})$. Moreover, it follows from the definition of $\mathcal{P}$ that $(\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}) \cap \mathrm{pa}(\mathcal{P}) = \mathrm{pa}(\mathcal{P}) \setminus \mathcal{P}$ and thus we have $\mathrm{pa}(\mathcal{O}) \setminus \mathcal{O} = (\mathrm{pa}(\mathcal{P}) \setminus \mathcal{P}) \cup (\mathrm{pa}(\mathcal{O}) \setminus (\mathcal{O} \cup \mathrm{pa}(\mathcal{P})))$. Now, pick an element $\hat{\boldsymbol{x}}_{\mathrm{pa}(\mathcal{O})\setminus(\mathcal{O}\cup\mathrm{pa}(\mathcal{P}))} \in \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus(\mathcal{O}\cup\mathrm{pa}(\mathcal{P}))}$ and define the mapping $\tilde{\boldsymbol{g}}_{\mathcal{P}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{P})\setminus\mathcal{P}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{P})} \to \boldsymbol{\mathcal{X}}_{\mathcal{P}}$ by

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

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

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

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

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

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

Section 4

Lemma 2.E.1. *Let $\mathcal{M}$ be an SCM that is uniquely solvable w.r.t. two subsets $A, B \subseteq \mathcal{I}$ that satisfy $A \subseteq B$ and $\mathrm{pa}(A) \setminus A \subseteq \mathrm{pa}(B) \setminus B$. Let $\boldsymbol{g}_A : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(A)\setminus A} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(A)} \to \boldsymbol{\mathcal{X}}_A$ and $\boldsymbol{g}_B : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(B)\setminus B} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(B)} \to \boldsymbol{\mathcal{X}}_B$ be measurable solution functions for $\mathcal{M}$ w.r.t. A and B, respectively. Then for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*

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

Proof. Without loss of generality, we assume that $\mathcal{M}$ is structurally minimal (see Proposition 2.2.11). Let $\bar{\boldsymbol{\mathcal{E}}} \subseteq \boldsymbol{\mathcal{E}}$ be a measurable set with $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\bar{\boldsymbol{\mathcal{E}}}) = 1$ such that for all $\boldsymbol{e} \in \bar{\boldsymbol{\mathcal{E}}}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$:

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

and

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

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

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

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

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

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

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

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

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

□

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

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

Proof of Proposition 2.4.6. First we show that equivalence implies counterfactual equivalence w.r.t. $\mathcal{O}$. The twin operation preserves the equivalence relation on SCMs and since equivalent SCMs are interventionally equivalent w.r.t. every subset, the two equivalent twin SCMs have to be interventionally equivalent w.r.t. $\mathcal{O} \cup \mathcal{O}'$ for every $\mathcal{O} \subseteq \mathcal{I}$ with $\mathcal{O}'$ the copy of $\mathcal{O}$ in $\mathcal{I}'$.

Now, let $\mathcal{M}$ and $\tilde{\mathcal{M}}$ be counterfactually equivalent w.r.t. $\mathcal{O}$. Then $\mathcal{M}^{\mathrm{twin}}$ and $\tilde{\mathcal{M}}^{\mathrm{twin}}$ are interventionally equivalent w.r.t. $\mathcal{O} \cup \mathcal{O}'$. Thus for $I \subseteq \mathcal{O}$, $I' \subseteq \mathcal{O}'$ the copy of I and $\boldsymbol{\xi}_{I'} = \boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, $\mathcal{M}^{\mathrm{twin}}_{\mathrm{do}(I\cup I', \boldsymbol{\xi}_{I\cup I'})}$ and $\tilde{\mathcal{M}}^{\mathrm{twin}}_{\mathrm{do}(I\cup I', \boldsymbol{\xi}_{I\cup I'})}$ are observationally equivalent w.r.t. $\mathcal{O} \cup \mathcal{O}'$. In particular, they are observationally equivalent w.r.t. $\mathcal{O}$. From Proposition 2.2.21 we have that $\mathcal{M}^{\mathrm{twin}}_{\mathrm{do}(I\cup I', \boldsymbol{\xi}_{I\cup I'})} = (\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)})^{\mathrm{twin}}$ and $\tilde{\mathcal{M}}^{\mathrm{twin}}_{\mathrm{do}(I\cup I', \boldsymbol{\xi}_{I\cup I'})} = (\tilde{\mathcal{M}}_{\mathrm{do}(I, \boldsymbol{\xi}_I)})^{\mathrm{twin}}$, and together with Lemma 2.E.2 this gives that $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}$ and $\tilde{\mathcal{M}}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}$ are observationally equivalent w.r.t. $\mathcal{O}$. □

Section 5

Lemma 2.E.3. *Let $\mathcal{M}$ be an SCM. Let $B \subseteq \mathcal{I}$ and $A \subseteq \mathcal{I} \cup \mathcal{J}$ such that $(\mathrm{pa}(B) \setminus B) \subseteq A$ and $B \cap A = \emptyset$. Assume that $g_B : \boldsymbol{\mathcal{X}}_A \times \boldsymbol{\mathcal{E}}_A \to \boldsymbol{\mathcal{X}}_B$ is a measurable function such that for $\mathbb{P}_{\mathcal{E}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*

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

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

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

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

Let $C := A \setminus (\mathrm{pa}(B) \setminus B)$, then by Lemma 2.F.11.(7) we have that there exists $\hat{\boldsymbol{e}}_C \in \boldsymbol{\mathcal{E}}_C$ and $\hat{\boldsymbol{x}}_C \in \boldsymbol{\mathcal{X}}_C$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}_{\mathcal{J}\setminus C}}$-almost every $\boldsymbol{e}_{\mathcal{J}\setminus C} \in \boldsymbol{\mathcal{E}}_{\mathcal{J}\setminus C}$ and for all $\boldsymbol{x}_{\mathcal{I}\setminus C} \in \boldsymbol{\mathcal{X}}_{\mathcal{I}\setminus C}$

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

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

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

where we picked $\hat{\boldsymbol{e}}_C \in \boldsymbol{\mathcal{E}}_C$ and $\hat{\boldsymbol{x}}_C \in \boldsymbol{\mathcal{X}}_C$ such that the above equivalence holds, and applying Lemma 2.F.11.(6) we get that for $\mathbb{P}_{\mathcal{E}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

Proof of Proposition 2.5.4. From unique solvability of $\mathcal{M}$ w.r.t. $\mathcal{L}_1$ it follows that there exists a mapping $\boldsymbol{g}_{\mathcal{L}_1} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{L}_1)\setminus(\mathcal{L}_1)} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{L}_1)} \to \boldsymbol{\mathcal{X}}_{\mathcal{L}_1}$ such that for $\mathbb{P}_{\mathcal{E}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

Let $\widehat{\mathrm{pa}}$ denotes the parents in $\mathcal{G}^a(\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)})$. Note that $\widehat{\mathrm{pa}}(\mathcal{L}_2) \setminus \mathcal{L}_2 \subseteq \mathrm{pa}(\mathcal{L}_1 \cup \mathcal{L}_2) \setminus (\mathcal{L}_1 \cup \mathcal{L}_2)$. Let $\tilde{f}$ denote the marginal causal mechanism of a structurally minimal SCM that is equivalent to the marginalization $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ constructed from $\boldsymbol{g}_{\mathcal{L}_1}$ (see Proposition 2.2.11).

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

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

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

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

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

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

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

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

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

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

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

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

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

Now apply Lemma 2.E.3 to conclude that $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$. □

Proof of Proposition 2.5.5. The commutation relation with the perfect intervention follows straightforwardly from the definitions of perfect intervention and marginalization and the fact that if $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$, then $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}$ is also uniquely solvable w.r.t. $\mathcal{L}$, since the structural equations for the variables $\mathcal{L}$ are the same for $\mathcal{M}$ and $\mathcal{M}_{\mathrm{do}(I, \boldsymbol{\xi}_I)}$.

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

Lemma 2.E.4. *Given an SCM $\mathcal{M}$ and a subset $\mathcal{L} \subseteq \mathcal{I}$ such that $\mathcal{M}$ is uniquely solvable w.r.t. $\mathcal{L}$. Then $\mathcal{M}$ and* $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ *are observationally equivalent w.r.t. $\mathcal{I} \setminus \mathcal{L}$.*

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

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

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

Proof of Theorem 2.5.6. The observational equivalence follows from Lemma 2.E.4. Using both Lemma 2.E.4 and Proposition 2.5.5 we can prove the interventional equivalence. Observe that from Proposition 2.5.5 we know that for a subset $I \subseteq \mathcal{I} \setminus \mathcal{L}$ and a value $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$, $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M})$ exists. By Lemma 2.E.4 we know that $\mathrm{do}(I, \boldsymbol{\xi}_I)(\mathcal{M})$ and $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I, \boldsymbol{\xi}_I))(\mathcal{M})$ are observationally equivalent w.r.t. $\mathcal{O}$ and hence by applying again Proposition 2.5.5, $\mathrm{do}(I, \boldsymbol{\xi}_I)(\mathcal{M})$ and $(\mathrm{do}(I, \boldsymbol{\xi}) \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$ are observationally equivalent w.r.t. $\mathcal{O}$. This implies that $\mathcal{M}$ and $\mathrm{marg}(\mathcal{L})(\mathcal{M})$ are interventionally equivalent w.r.t. $\mathcal{O}$. Lastly, we need to show that $\mathrm{twin}(\mathcal{M})$ and $(\mathrm{twin} \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$ are interventionally equivalent w.r.t. $(\mathcal{I} \cup \mathcal{I}') \setminus (\mathcal{L} \cup \mathcal{L}')$, where $\mathcal{L}'$ is the copy of $\mathcal{L}$ in $\mathcal{I}'$. From Proposition 2.5.5 $(\mathrm{twin} \circ \mathrm{marg}(\mathcal{L}))(\mathcal{M})$ is equivalent to $(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin})(\mathcal{M})$ and since we proved that $(\mathrm{marg}(\mathcal{L} \cup \mathcal{L}') \circ \mathrm{twin})(\mathcal{M})$ and $\mathrm{twin}(\mathcal{M})$ are interventionally equivalent w.r.t. $(\mathcal{I} \cup \mathcal{I}') \setminus (\mathcal{L} \cup \mathcal{L}')$ the result follows. □

Proof of Proposition 2.5.8. A similar proof as for Theorem 1 in (Evans, 2016) works. □

Proof of Proposition 2.5.9. First we prove the commutation relation of the perfect intervention. Observe that applying the $\mathrm{do}(I)$ operation to the latent projection $\mathrm{marg}(\mathcal{L})(\mathcal{G})$ removes all the incoming edges on the nodes I. Such an incoming edge at a node in I in $\mathrm{marg}(\mathcal{L})(\mathcal{G})$ corresponds to a path in $\mathcal{G}$ that points to that node. But since $\mathrm{do}(I)(\mathcal{G})$ is just $\mathcal{G}$ with all the incoming edges on I removed, the graph $(\mathrm{marg}(\mathcal{L}) \circ \mathrm{do}(I))(\mathcal{G})$ also has all the incoming edges on the nodes I removed.

Next, we will prove the commutation relation of the twin operation. We will denote the copy in $\mathcal{I}'$ of any node $i \in \mathcal{I}$ by i', that is, $\mathcal{I}' = \{i' : i \in \mathcal{I}\}$. The edges in $(\text{twin}(\mathcal{I} \setminus \mathcal{L}) \circ \text{marg}(\mathcal{L}))(\mathcal{G})$ can be partitioned into three cases:

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

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

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

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

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

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

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

□

Proof of Proposition 2.5.11. Without loss of generality, we assume that $\mathcal{M}$ is structurally minimal (see Proposition 2.2.11). Let $\boldsymbol{g}_{\mathcal{L}}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{L}$ and denote by $\mathcal{M}_{\text{marg}(\mathcal{L})}$ the marginal SCM constructed from $\boldsymbol{g}_{\mathcal{L}}$. For $j \in \mathcal{I} \setminus \mathcal{L}$, define $A_j := \text{an}_{\mathcal{G}(\mathcal{M})_{\mathcal{L}}}(\text{pa}(j) \cap \mathcal{L}) \subseteq \mathcal{L}$ and let $\tilde{\boldsymbol{g}}_{A_j}$ be a measurable solution function for $\mathcal{M}$ w.r.t. A_j. Because $A_j \subseteq \mathcal{L}$ and $\text{pa}(A_j) \setminus A_j \subseteq \text{pa}(\mathcal{L}) \setminus \mathcal{L}$, by Lemma 2.E.1, for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

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

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

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

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

Section 6

Proof of Theorem 2.6.3. This follows directly from Theorems 2.A.7 and 2.A.21. □

Section 7

Proof of Proposition 2.7.1. We define $\tilde{\mathcal{M}} := \mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$, $\widetilde{\mathrm{pa}} := \mathrm{pa}_{\mathcal{G}^a(\tilde{\mathcal{M}})}$ and $\mathcal{A} := \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})_{\setminus i}}(j)$. Suppose that $i \to j \notin \mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$ and assume that the two induced distributions do not coincide. Because $i \to j \notin \mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$ it follows that $(\widetilde{\mathrm{pa}}(\mathcal{A}) \setminus \mathcal{A}) \cap \mathcal{I} = \emptyset$. Let now $\tilde{\boldsymbol{g}}_{\mathcal{A}} : \boldsymbol{\mathcal{E}}_{\widetilde{\mathrm{pa}}(\mathcal{A})} \to \boldsymbol{\mathcal{X}}_{\mathcal{A}}$ be a measurable solution function for $\tilde{\mathcal{M}}$ w.r.t. $\mathcal{A}$, that is, we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

Proof of Proposition 2.7.2. Let's define $\tilde{\mathcal{M}} := \mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$, $\widetilde{\mathrm{pa}} := \mathrm{pa}_{\mathcal{G}^a(\tilde{\mathcal{M}})}$, $\mathcal{A}_i := \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})}(i)$ and $\mathcal{A}_j^{\setminus i} := \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})_{\setminus i}}(j)$. Suppose that there does not exist a bidirected edge $i \leftrightarrow j$ in the latent projection $\mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\mathcal{M}))$. Because $i \leftrightarrow j \notin \mathrm{marg}(\mathcal{I} \setminus \mathcal{O})(\mathcal{G}(\tilde{\mathcal{M}}))$, where here $\tilde{\mathcal{M}}$ is the intervened model $\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$, we have that $\mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus j}}(i) \cap$

$\mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus i}}(j) \cap \mathcal{J} = \emptyset$. From $j \notin \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})}(i)$ it follows that $\mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})_{\setminus j}}(i) = \mathrm{an}_{\mathcal{G}(\tilde{\mathcal{M}})}(i)$, and hence $\mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})}(i) \cap \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus i}}(j) \cap \mathcal{J} = \emptyset$. Observe that $\widetilde{\mathrm{pa}}(\mathcal{A}_i) \subseteq \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})}(i)$ and $\widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \subseteq \mathrm{an}_{\mathcal{G}^a(\tilde{\mathcal{M}})_{\setminus i}}(j) \cup \{i\}$, and thus $\widetilde{\mathrm{pa}}(\mathcal{A}_i) \cap \widetilde{\mathrm{pa}}(\mathcal{A}_j^{\setminus i}) \cap \mathcal{J} = \emptyset$. Let $\boldsymbol{g}_{\mathcal{A}_i} : \boldsymbol{\mathcal{E}}_{\widetilde{\mathrm{pa}}(\mathcal{A}_i)} \to \boldsymbol{\mathcal{X}}_{\mathcal{A}_i}$ be a measurable solution function for $\tilde{\mathcal{M}}$ w.r.t. $\mathcal{A}_i$, that is, we have for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

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

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

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

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

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

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

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

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

Section 8

Proof of Proposition 2.8.2. We first show that the class of simple SCMs is closed under marginalization. Take two disjoint subsets $\mathcal{L}_1$ and $\mathcal{L}_2$ in $\mathcal{I}$. Then, it suffices to show that $\mathcal{M}_{\mathrm{marg}(\mathcal{L}_1)}$ is uniquely solvable w.r.t. $\mathcal{L}_2$. This follows directly from Proposition 2.5.4.

To show that the class of simple SCMs is closed under perfect intervention. Let $\mathcal{M}$ be a simple SCM, $\mathcal{O} \subseteq \mathcal{I}$, $I \subseteq \mathcal{I}$ and $\boldsymbol{\xi}_I \in \boldsymbol{\mathcal{X}}_I$. Define $\mathcal{O}_1 := \mathcal{O} \cap I$ and $\mathcal{O}_2 := \mathcal{O} \setminus I$,

then $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2$. Note that $\mathrm{pa}(\mathcal{O}_2) \setminus \mathcal{O}_2 = (\mathrm{pa}(\mathcal{O}_2) \setminus (\mathcal{O}_2 \cup I)) \cup (\mathrm{pa}(\mathcal{O}_2) \cap I)$ and $\mathrm{pa}(\mathcal{O}_2) \setminus (\mathcal{O}_2 \cup I) \subseteq \mathrm{pa}(\mathcal{O}) \setminus \mathcal{O}$. Let $\boldsymbol{g}_{\mathcal{O}_2} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O}_2)\setminus\mathcal{O}_2} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O}_2)} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}_2}$ be a measurable solution function for $\mathcal{M}$ w.r.t. $\mathcal{O}_2$. The mapping $\tilde{\boldsymbol{g}}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{\mathrm{pa}(\mathcal{O})\setminus\mathcal{O}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}(\mathcal{O})} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ defined by

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

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

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

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

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

Lastly, it follows that the observational and all the intervened models of $\mathcal{M}$ and $\mathcal{M}^{\mathrm{twin}}$ are uniquely solvable. From Theorem 2.3.6 we conclude that $\mathcal{M}$ induces unique observational, interventional and counterfactual distributions. □

Proof of Corollary 2.8.3. This follows from Corollary 2.A.22. □

2.F MEASURABLE SELECTION THEOREMS

In this appendix, we derive some lemmas and state two measurable selection theorems that are used in several proofs in Appendix 2.E. First, we introduce the measure theoretic notation and terminology needed to understand the results (see (Kechris, 1995) for more details).

Definition 2.F.1 (Standard measurable space)**.** *A measurable space $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Sigma})$ is a* standard measurable space *if it is isomorphic to $(\boldsymbol{\mathcal{Y}}, \mathcal{B}(\boldsymbol{\mathcal{Y}}))$, where $\boldsymbol{\mathcal{Y}}$ is a Polish space, that is, a separable completely metrizable space,[24] and $\mathcal{B}(\boldsymbol{\mathcal{Y}})$ are the Borel subsets of $\boldsymbol{\mathcal{Y}}$, that is,*

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

the σ-algebra generated by the open sets in $\mathcal{Y}$. A measure space $(\mathcal{X}, \Sigma, \mu)$ is a standard probability space *if $(\mathcal{X}, \Sigma)$ is a standard measurable space and μ is a probability measure.*

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

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

Lemma 2.F.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be standard measurable spaces and $f : \mathcal{X} \to \mathcal{Y}$ a measurable mapping. Then*

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

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

Definition 2.F.4 (μ-measurability)**.** *Let $(\mathcal{X}, \Sigma, \mu)$ be a measure space. A set $\mathcal{E} \subseteq \mathcal{X}$ is called a μ-*null set *if there exists a $\mathcal{A} \in \Sigma$ with $\mathcal{E} \subseteq \mathcal{A}$ and $\mu(\mathcal{A}) = 0$. We denote the class of μ-null sets by $\mathcal{N}$, and we denote the σ-algebra generated by $\Sigma \cup \mathcal{N}$ by $\bar{\Sigma}$, and its members are called the μ-*measurable *sets. Note that each member of $\bar{\Sigma}$ is of the form $\mathcal{A} \cup \mathcal{E}$ with $\mathcal{A} \in \Sigma$ and $\mathcal{E} \in \mathcal{N}$. The measure μ is extended to a measure $\bar{\mu}$ on $\bar{\Sigma}$, by $\bar{\mu}(\mathcal{A} \cup \mathcal{E}) = \mu(\mathcal{A})$ for every $\mathcal{A} \in \Sigma$ and $\mathcal{E} \in \mathcal{N}$, and is called its* completion*. A mapping $f : \mathcal{X} \to \mathcal{Y}$ between measurable spaces is called μ-*measurable *if the inverse image $f^{-1}(\mathcal{C})$ of every measurable set $\mathcal{C} \subseteq \mathcal{Y}$ is μ-measurable.*

Definition 2.F.5 (Universal measurability)**.** *Let $(\mathcal{X}, \Sigma)$ be a standard measurable space. A set $\mathcal{A} \subseteq \mathcal{X}$ is called* universally measurable *if it is μ-measurable for every σ-finite measure[25] μ on $\mathcal{X}$ (i.e., in particular every probability measure). A mapping $f : \mathcal{X} \to \mathcal{Y}$ between standard measurable spaces is* universally measurable *if it is μ-measurable for every σ-finite measure μ.*

Lemma 2.F.6. *Let $\mathcal{E}$ be a standard probability space with probability measure $\mathbb{P}_{\mathcal{E}}$ and $\mathcal{A} \subseteq \mathcal{E}$ an analytic set. Then $\mathcal{A}$ is $\mathbb{P}_{\mathcal{E}}$-measurable and there exist measurable sets $\mathcal{S}, \mathcal{T} \subseteq \mathcal{E}$ such that $\mathcal{S} \subseteq \mathcal{A} \subseteq \mathcal{T}$ and $\mathbb{P}_{\mathcal{E}}(\mathcal{S}) = \bar{\mathbb{P}}_{\mathcal{E}}(\mathcal{A}) = \mathbb{P}_{\mathcal{E}}(\mathcal{T})$, where $\bar{\mathbb{P}}_{\mathcal{E}}$ is the completion of $\mathbb{P}_{\mathcal{E}}$.*

---

separable completely metrizable space is called a *Polish space* (see (Cohn, 2013) and (Kechris, 1995) for more details).

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

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

Lemma 2.F.7. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a μ-measurable mapping. If $\mathcal{Y}$ is countably generated, then there exists a measurable mapping $g : \mathcal{X} \to \mathcal{Y}$ such that $f(x) = g(x)$ holds μ-a.e..*

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

$$g(x) := \begin{cases} f(x) & \text{if } x \in \hat{\mathcal{A}}, \\ y_0 & \text{otherwise,} \end{cases}$$

where for y_0 we can take an arbitrary point in $\mathcal{Y}$. This mapping g is measurable since for each generator $\mathcal{C}_n$ we have

$$g^{-1}(\mathcal{C}_n) = \begin{cases} \hat{\mathcal{A}}_n & \text{if } y_0 \notin \mathcal{C}_n, \\ \hat{\mathcal{A}}_n \cup \hat{\mathcal{B}} & \text{otherwise.} \end{cases}$$

is in Σ. Moreover, $f(x) = g(x)$ μ-almost everywhere. □

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

Theorem 2.F.8 (Measurable selection theorem)**.** *Let $\mathcal{E}$ be a standard probability space with probability measure $\mathbb{P}_{\mathcal{E}}$, $\mathcal{X}$ a standard measurable space and $\mathcal{S} \subseteq \mathcal{E} \times \mathcal{X}$ a measurable set such that $\mathcal{E} \setminus pr_{\mathcal{E}}(\mathcal{S})$ is a $\mathbb{P}_{\mathcal{E}}$-null set, where $pr_{\mathcal{E}} : \mathcal{E} \times \mathcal{X} \to \mathcal{E}$ is the projection mapping on $\mathcal{E}$. Then there exists a measurable mapping $g : \mathcal{E} \to \mathcal{X}$ such that $(e, g(e)) \in \mathcal{S}$ for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$.*

Proof. Take the subset $\hat{\mathcal{E}} := \mathcal{E} \setminus \mathcal{B}$, for some measurable set $\mathcal{B} \supseteq \mathcal{E} \setminus pr_{\mathcal{E}}(\mathcal{S})$ and $\mathbb{P}_{\mathcal{E}}(\mathcal{B}) = 0$, and note that $\hat{\mathcal{E}}$ is a standard measurable space (see Corollary 13.4 in (Kechris, 1995)) and $\hat{\mathcal{E}} \subseteq pr_{\mathcal{E}}(\mathcal{S})$. Let $\hat{\mathcal{S}} = \mathcal{S} \cap (\hat{\mathcal{E}} \times \mathcal{X})$. Because the set $\hat{\mathcal{S}}$ is measurable, it is in particular analytic (see Lemma 2.F.3). It follows by the Jankov-von Neumann Theorem (see Theorem 18.8 or 29.9 in (Kechris, 1995)) that $\hat{\mathcal{S}}$ has a universally measurable uniformizing function, that is, there exists a universally measurable mapping $\hat{g} : \hat{\mathcal{E}} \to \mathcal{X}$ such that for all $e \in \hat{\mathcal{E}}$, $(e, \hat{g}(e)) \in \hat{\mathcal{S}}$. Hence, in particular, it is $\mathbb{P}_{\mathcal{E}}\big|_{\hat{\mathcal{E}}}$-measurable, where $\mathbb{P}_{\mathcal{E}}\big|_{\hat{\mathcal{E}}}$ is the restriction of $\mathbb{P}_{\mathcal{E}}$ to $\hat{\mathcal{E}}$.

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

$$g^*(e) := \begin{cases} \hat{g}(e) & \text{if } e \in \hat{\mathcal{E}} \\ x_0 & \text{otherwise,} \end{cases}$$

where for x_0 we can take an arbitrary point in $\mathcal{X}$. Then this mapping g^* is $\mathbb{P}_{\mathcal{E}}$-measurable. To see this, take any measurable set $\mathcal{C} \subseteq \mathcal{X}$, then

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

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

By Lemma 2.F.7 and the fact that standard measurable spaces are countably generated (see Proposition 12.1 in (Kechris, 1995)), we prove the existence of a measurable mapping $g : \mathcal{E} \to \mathcal{X}$ such that $g^* = g$ $\mathbb{P}_{\mathcal{E}}$-a.e. and thus it satisfies $(e, g(e)) \in \mathcal{S}$ for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$. □

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

Theorem 2.F.9 (Second measurable selection theorem)**.** *Let $\mathcal{E}$ be a standard probability space with probability measure $\mathbb{P}_{\mathcal{E}}$, $\mathcal{X}$ and $\mathcal{Y}$ standard measurable spaces and $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{E} \times \mathcal{Y}$ a measurable set such that $\mathcal{E} \setminus \mathcal{K}_\sigma$ is a $\mathbb{P}_{\mathcal{E}}$-null set, where*

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

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

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

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

Proof. Take the subset $\hat{\mathcal{E}} := \mathcal{E} \setminus \mathcal{B}$, for some measurable set $\mathcal{B} \supseteq \mathcal{E} \setminus \mathcal{K}_\sigma$ and $\mathbb{P}_{\mathcal{E}}(\mathcal{B}) = 0$. Note that $\hat{\mathcal{E}}$ is a standard measurable space, $\hat{\mathcal{E}} \subseteq \mathcal{K}_\sigma$ and $\hat{\mathcal{S}} = \mathcal{S} \cap (\mathcal{X} \times \hat{\mathcal{E}} \times \mathcal{Y})$ is measurable. By assumption, for each $(x, e) \in \mathcal{X} \times \hat{\mathcal{E}}$ the fiber $\hat{\mathcal{S}}_{(x,e)}$ is nonempty and σ-compact and hence by applying the Theorem of Arsenin-Kunugui (see Theorem 35.46 in (Kechris, 1995)) it follows that the set $\hat{\mathcal{S}}$

has a measurable uniformizing function, that is, there exists a measurable mapping $\hat{g} : \boldsymbol{\mathcal{X}} \times \hat{\boldsymbol{\mathcal{E}}} \to \boldsymbol{\mathcal{Y}}$ such that for all $(\boldsymbol{x}, \boldsymbol{e}) \in \boldsymbol{\mathcal{X}} \times \hat{\boldsymbol{\mathcal{E}}}$, $(\boldsymbol{x}, \boldsymbol{e}, \hat{g}(\boldsymbol{x}, \boldsymbol{e})) \in \hat{\boldsymbol{\mathcal{S}}}$. Now define the mapping $g : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{Y}}$ by

$$g(\boldsymbol{x}, \boldsymbol{e}) := \begin{cases} \hat{g}(\boldsymbol{x}, \boldsymbol{e}) & \text{if } \boldsymbol{e} \in \hat{\boldsymbol{\mathcal{E}}} \\ y_0 & \text{otherwise,} \end{cases}$$

where for y_0 we can take an arbitrary point in $\boldsymbol{\mathcal{Y}}$. This mapping g inherits the measurability from $\hat{g}$ and it satisfies for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ that $(\boldsymbol{x}, \boldsymbol{e}, g(\boldsymbol{x}, \boldsymbol{e})) \in \boldsymbol{\mathcal{S}}$. □

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

Lemma 2.F.10. *Let $\boldsymbol{\phi} : \boldsymbol{\mathcal{E}} \to \tilde{\boldsymbol{\mathcal{E}}}$ be a measurable map between two standard measurable spaces. Let $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$ be a probability measure on $\boldsymbol{\mathcal{E}}$ and let $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}} \circ \boldsymbol{\phi}^{-1}$ be its push-forward under $\boldsymbol{\phi}$. Let $\tilde{P} : \tilde{\boldsymbol{\mathcal{E}}} \to \{0, 1\}$ be a property, that is, a (measurable) boolean-valued function on $\tilde{\boldsymbol{\mathcal{E}}}$. Then the property $P = \tilde{P} \circ \boldsymbol{\phi}$ on $\boldsymbol{\mathcal{E}}$ holds $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.e. if and only if the property $\tilde{P}$ holds $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-a.e..*

Proof. Assume the property $P = \tilde{P} \circ \boldsymbol{\phi}$ holds $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-a.e., then $\boldsymbol{\mathcal{C}} = \{\boldsymbol{e} \in \boldsymbol{\mathcal{E}} : P(\boldsymbol{e}) = 1\}$ contains a measurable set $\boldsymbol{\mathcal{C}}^*$ with $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-measure 1, that is, $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\mathcal{C}}$ and $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{C}}^*) = 1$. By Lemma 2.F.3, $\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*)$ is analytic. By Lemma 2.F.6, there exist measurable sets $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}$ such that $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*) \subseteq \boldsymbol{\mathcal{B}}$ and $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{A}}) = \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{B}})$. Because $\boldsymbol{\phi}$ is measurable, $\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{A}})$ and $\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{B}})$ are both measurable. Also, $\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{A}}) \subseteq \boldsymbol{\phi}^{-1}(\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*)) \subseteq \boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{B}})$. As $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\phi}^{-1}(\boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*))$, we must have that $\mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\phi}^{-1}(\boldsymbol{\mathcal{B}})) \geq \mathbb{P}_{\boldsymbol{\mathcal{E}}}(\boldsymbol{\mathcal{C}}^*) = 1$. Hence $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{A}}) = \mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}(\boldsymbol{\mathcal{B}}) = 1$. Note that as $\boldsymbol{\mathcal{C}}^* \subseteq \boldsymbol{\mathcal{C}}$, $\boldsymbol{\mathcal{A}} \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}^*) \subseteq \boldsymbol{\phi}(\boldsymbol{\mathcal{C}}) \subseteq \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}}) = 1\}$. Hence the set $\tilde{\boldsymbol{\mathcal{C}}} := \{\tilde{\boldsymbol{e}} \in \tilde{\boldsymbol{\mathcal{E}}} : \tilde{P}(\tilde{\boldsymbol{e}}) = 1\}$ contains a measurable set of $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-measure 1, in other words, $\tilde{P}$ holds $\mathbb{P}_{\tilde{\boldsymbol{\mathcal{E}}}}$-a.s..

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

Lemma 2.F.11 (Some properties for the for-almost-every quantifier)**.** *Let $\boldsymbol{\mathcal{X}} = \mathcal{X} \times \tilde{\mathcal{X}}$ and $\boldsymbol{\mathcal{E}} = \mathcal{E} \times \tilde{\mathcal{E}}$ be products of nonempty standard measurable spaces and $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \mathbb{P}_{\mathcal{E}} \times \mathbb{P}_{\tilde{\mathcal{E}}}$ be the product measure of probability measures $\mathbb{P}_{\mathcal{E}}$ and $\mathbb{P}_{\tilde{\mathcal{E}}}$ on $\mathcal{E}$ and $\tilde{\mathcal{E}}$, respectively. Denote by "$\forall \boldsymbol{e}$" the quantifier "for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" and by "$\forall \boldsymbol{x}$" the quantifier "for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$", and similarly for their components, for example, "$\forall e$" for "for $\mathbb{P}_{\mathcal{E}}$-almost every $e \in \mathcal{E}$" and "$\forall x$" for "for all $x \in \mathcal{X}$". Then we have the following properties:*

1. $\forall e : P(e) \implies \exists e : P(e)$
   (similarly to $\forall x : P(x) \implies \exists x : P(x)$);
2. $\forall e : P(e) \iff \forall \boldsymbol{e} : P(e)$
   (similarly to $\forall x : P(x) \iff \forall \boldsymbol{x} : P(x)$);

3. $\exists x \forall^* e : P(x,e) \implies \forall^* e \exists x : P(x,e)$
   (similarly to $\exists x \forall e : P(x,e) \implies \forall e \exists x : P(x,e)$*);*

4. $\forall^* e \forall x : P(x,e) \implies \forall x \forall^* e : P(x,e)$
   (similarly to $\forall e \forall x : P(x,e) \implies \forall x \forall e : P(x,e)$*);*

5. $\forall^* \boldsymbol{e} : P(\boldsymbol{e}) \implies \exists \tilde{e} \forall^* e : P(\boldsymbol{e})$
   (similarly to $\forall \boldsymbol{x} : P(\boldsymbol{x}) \implies \exists \tilde{x} \forall x : P(\boldsymbol{x})$*);*

6. $\forall^* e \forall x : P(x,e) \iff \forall^* \boldsymbol{e} \forall \boldsymbol{x} : P(x,e)$*;*

7. $\forall^* \boldsymbol{e} \forall \boldsymbol{x} : P(\boldsymbol{x},\boldsymbol{e}) \implies \exists \tilde{e} \exists \tilde{x} \forall^* e \forall x : P(\boldsymbol{x},\boldsymbol{e})$*,*

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

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

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

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

Property 4: We have

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

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

Property 7: We have

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

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

3

CAUSAL MODELING OF DYNAMICAL SYSTEMS

Dynamical systems are widely used in science and engineering to model systems consisting of several interacting components. Often, they can be given a *causal* interpretation in the sense that they not only model the evolution of the states of the system's components over time, but also describe how their evolution is affected by external interventions on the system that perturb the dynamics. We introduce the formal framework of structural dynamical causal models (SDCMs) that explicates the causal semantics of the system's components as part of the model. SDCMs represent a dynamical system as a collection of stochastic processes and specify the basic causal mechanisms that govern the dynamics of each component as a structured system of random differential equations of arbitrary order. SDCMs extend the versatile causal modeling framework of structural causal models (SCMs), also known as structural equation models (SEMs), by explicitly allowing for time-dependence. An SDCM can be thought of as the stochastic-process version of an SCM, where the static random variables of the SCM are replaced by dynamic stochastic processes and their derivatives. We provide the foundations for a theory of SDCMs, by (i) formally defining SDCMs, their solutions, stochastic interventions, and a graphical representation; (ii) studying existence and uniqueness of the solutions for given initial conditions; (iii) providing Markov properties for SDCMs with initial conditions; (iv) discussing under which conditions SDCMs equilibrate to SCMs as time tends to infinity; (v) relating the properties of the SDCM to those of the equilibrium SCM. This correspondence enables one to leverage the wealth of statistical tools and discovery methods available for SCMs when studying the causal semantics of a large class of stochastic dynamical systems. The theory is illustrated with several well-known examples from different scientific domains.[1]

3.1 INTRODUCTION

Continuous dynamical systems consisting of differential equations are widely used in science and engineering to model the time-dependent behavior of certain phenomena. A classical example is the modeling of the trajectory of a die that is thrown, by means of Newton's equations of motion. Initial conditions or parameters of the dynamics may be stochastic, which can be modeled mathematically by making use of random differential equations (RDEs). These provide a natural extension of ordinary differential equations (ODEs) to the stochastic setting (Bunke, 1972; Soong, 1973; Sobczyk, 1991; Neckel and Rupp, 2013). For example, the initial position of the die is often not known, and varies from throw to throw, which leads

1 The material in this chapter has been submitted to the Journal of Causal Inference. A preprint is available as (Bongers, Blom, and Mooij, 2022).

to a probability distribution over the possible trajectories of the die (and eventually, to an uncertain outcome of the throw).

Many dynamical systems can be considered to consist of several interacting subsystems or components, for example, mass-spring systems in physics, predator-prey systems in biology, and mass-action law kinetics in chemistry. These dynamical systems are often implicitly given a *causal* interpretation in the sense that they are not only supposed to model the evolution of the state of the system over time, but also describe how the evolution of the system's components is affected by external interventions on the system that perturb the dynamics. For example, when applying an external force to a particle, the change in the force term in Newton's second law of motion results in a changed acceleration, and hence a changed position, of the particle. Another example is that hunting wolves may lead to an increase in the population of sheep. The ensuing causal semantics of the system is usually only treated in an implicit and intuitive fashion, rather than that it is formally specified by (or derivable from) the mathematical model. Indeed, a system of (random) differential equations simply expresses symmetric relations between the components, without any preferred order or asymmetry. On the other hand, causal relations may be asymmetric, as they distinguish cause from effect. Thus, while dynamical systems may describe how the state of a system consisting of several components evolves over time, by themselves they do not express the inherent "causal structure" of the system's components.

An apparently rather different modeling framework that allows to represent the causal semantics of a system composed of components is provided by structural causal models (SCMs), also known as (non-parametric) structural equation models (SEMs) (Bollen, 1989; Spirtes, Glymour, and Scheines, 2000; Peters, Janzing, and Schölkopf, 2017; Bongers et al., 2021). First introduced in genetics by Wright (1921), they became popular over the years in econometrics (Haavelmo, 1943), the social sciences (Goldberger and Duncan, 1973; Duncan, 1975), and more recently in AI (Pearl, 2009). SCMs express causal relationships between variables corresponding to "autonomous" subsystems or components in the form of deterministic, functional relationships, and stochasticity is introduced through the assumption that certain variables are exogenous (latent) random variables. Their predictive power stems from the assumption that the equations of these models are organized in a structural way: each equation represents a distinct autonomous causal mechanism, where distinctness of the mechanisms means that they can be changed independently of one another by targeted interventions—at least in principle. While SCMs explicate the causal semantics of a system composed of different components in this specific way, they have no built-in notion of time. A commonly used workaround for this limitation is to introduce multiple "copies" of the variables, corresponding to observations at different points (or intervals) in time. This workaround only applies to discrete time, and SCMs cannot be used to model causal semantics of continuous-time systems without somehow discretizing time.

In this work, we propose the modeling framework of *structural dynamical causal models* (SDCMs), which on the one hand explicates the causal relationships between

components of continuous dynamical systems, and on the other hand extends structural causal models to explicitly allow for time-dependence. SDCMs represent a dynamical system as a collection of stochastic processes (each one referring to a causally "autonomous" component) subject to a "structured" dynamics, which specifies the causal mechanisms that govern the dynamics of the components by means of random differential equations of arbitrary order. An SDCM can be thought of as the stochastic-process version of an SCM, where the static (time-independent) random variables of the SCM are replaced by dynamic (time-dependent) stochastic processes and their derivatives. Our contributions can be considered as the first steps towards a theory of SDCMs. More specifically, we:

(i) formally define SDCMs, their solutions, stochastic interventions, and a graphical representation;

(ii) study existence and uniqueness of the solutions for given initial conditions;

(iii) provide Markov properties for SDCMs with initial conditions;

(iv) discuss under which conditions SDCMs equilibrate to SCMs as time tends to infinity;

(v) relate the properties of the SDCM to those of the equilibrium SCM.

This correspondence between SCMs and equilibrated SDCMs enables one to leverage the wealth of statistical tools and discovery methods available for SCMs when studying the causal semantics of a large class of stochastic dynamical systems. We illustrate the theory with several well-known examples from different scientific domains.

RELATED WORK Over the years, several efforts have been made to develop a notion of causality for stochastic processes, both in discrete and continuous time.

For discrete time, Granger causality (Granger, 1969; White, 2006; Eichler, 2007; Eichler and Didelez, 2007), simultaneous equation models (Fisher, 1970; Lacerda et al., 2008), vector autoregressive (VAR) models (Sims, 1980; Lütkepohl, 2005) and dynamic Bayesian networks (Dagum, Galper, and Horvitz, 1992; Ghahramani, 1998) have been studied extensively. More recently, there has been some work on learning difference-based causal models (Voortman, Dash, and Druzdzel, 2010) and structural equation models (Peters, Janzing, and Schölkopf, 2013). In principle, all these models fit directly into the framework of SCMs by labeling the random variables with time.

For continuous time, there has been substantial work in the graphical modeling community (Aalen, 1987; Didelez, 2000, 2007, 2008, 2015) based on the concept of local independence, which was introduced by Schweder (1970). However, none of these approaches explicitly takes into account that dynamical models are often based on differential equations. In parallel, several attempts have been made to arrive at causal interpretations of processes described by ordinary and stochastic differential equations. Many of these approaches start from the assumption of a

first-order system of ODEs written in canonical form, and implicitly (or explicitly) attribute a causal interpretation to this (Iwasaki and Simon, 1994; Mooij, Janzing, and Schölkopf, 2013; Pfister, Bauer, and Peters, 2019; Blom and Mooij, 2021). The notion of causality in ODEs has also been studied using Simon's causal ordering algorithm (Iwasaki and Simon, 1994). Relations between a certain class of causally interpreted ODEs and deterministic SCMs at equilibrium have been established under the strong assumption that all the solutions of the ODE converge to a single static equilibrium state (Mooij, Janzing, and Schölkopf, 2013), independent of the initial condition. This assumption can be relaxed to allow for asymptotic dynamics (Rubenstein et al., 2018) such as periodic oscillations, but this still requires the assumption that the asymptotic dynamics does not depend on the initial condition. Another way to relax the assumption of (Mooij, Janzing, and Schölkopf, 2013) is taken in the framework of causal constraints models (Blom, Bongers, and Mooij, 2019), which can model static equilibrium states as long as the dynamical system has a unique static equilibrium state corresponding to each initial condition, for every intervention. These models can give a more complete causal description of these static equilibrium states than SCMs can (Blom, Bongers, and Mooij, 2019), but this comes at the cost that they appear to be too "flexible" in general. Finally, several approaches in terms of stochastic differential equations, which are differential equations with an additive white noise term, have been developed over the years (Florens and Fougere, 1996; Commenges and Gégout-Petit, 2009; Hansen and Sokol, 2014; Mogensen, Malinsky, and Hansen, 2018; Peters, Bauer, and Pfister, 2020). The stochastic differential equations have the advantage that they can deal with "instantaneous" stochasticity in the dynamics, but solving them usually requires a considerable mathematical effort using Itô calculus.

Compared with existing work, the framework of structural dynamical causal models that we propose here has the novel combination of features that it extends the semantics of continuous dynamical systems by formally encoding the causal structure into the model, it allows for stochasticity due to uncertainty over initial conditions or parameters of the dynamics without relying on strong stability assumptions, and it does not force one to consider time derivatives of processes as being "causally independent" of the processes themselves (that is, time derivatives of processes are considered to describe the same subsystem or component as the process itself). Our framework reconciles the traditional intuitive treatment of causality in the context of deterministic dynamical systems as practiced in many exact sciences with the treatment of causality of stochastic systems that is nowadays very popular in AI, statistics and other scientific disciplines. An attractive feature is that it naturally accommodates many causally interpreted continuous dynamical systems that appear "in the wild".

CONTRIBUTIONS In this chapter, we introduce the framework of *structural dynamical causal models* (SDCMs),[2] which allows to model the causal semantics of

2 Not to be confused with the *dynamic causal models* of (Friston, Harrison, and Penny, 2003) or the *dynamic structural causal models* of (Rubenstein et al., 2018). The dynamic causal models of (Friston,

stochastic processes for a large class of continuous dynamical systems by means of a "structured" system of random differential equations of arbitrary order (including zeroth-order). One can consider SCMs as special cases of SDCMs that only contain zeroth-order equations. The proposed modeling framework enables modeling of stochasticity, time-dependence and causality in a natural way. We study the existence and uniqueness of solutions of SDCMs, and propose a convenient graphical representation of the model structure for which we derive Markov properties. We define an idealized notion of stochastic interventions, and show that this yields a natural "interventionist" causal interpretation of the graph of an SDCM. We define a notion of equilibration of an SDCM to an SCM, which corresponds with letting a system converge towards equilibrium as time tends to infinity, and relate the properties of the SDCM to those of the equilibrium SCM. In the next paragraphs, we describe our contributions in more detail.

Intuitively, an SDCM can be thought of as an SCM where the notion of time is added to the structural equations by replacing the random variables of the SCM by stochastic processes and their (higher-order) derivatives. In the presence of these derivative processes, these equations, which we coin *dynamic structural equations*, can be read as random differential equations. The dynamic structural equations have the property that they are organized in a structural way, similar to how the structural equations of an SCM are organized by associating a distinct causal mechanism to each observed variable. This distinguishes SDCMs from other "non-causal" (random) dynamical systems, and allows to define idealized stochastic interventions on these models, similarly to how this is usually done for SCMs. The structure of the SCM can be expressed by its graph, which reflects the functional relationships between the components as encoded by the structural equations. Similarly, we define the graph of an SDCM to reflect the functional relationships between the components as encoded by the dynamic structural equations.

The framework of SDCMs on the one hand allows one to specify the causal semantics of a system of RDEs, and on the other hand it enables temporal extensions for SCMs. In particular, we show when and how we can equilibrate an SDCM to an SCM, such that the static solutions of the SCM contain the equilibrium states of the SDCM. Our equilibration operation, inspired by the one of Mooij, Janzing, and Schölkopf (2013), has the key property that it preserves the structure of the endogenous processes. Intuitively, the idea is that in the limit as time tends to infinity, the dynamic structural equations converge to those equations for which the higher-order derivatives of the processes have been set to zero, yielding the structural equations of an SCM. This allows us to use SCMs to model the equilibrium states of dynamical systems, including cases that were previously considered to fall outside their scope, such as the price, supply and demand model in econometrics. In

Harrison, and Penny, 2003) have been developed to infer the causal relations between the activities of different brain regions, where each neuronal state is modeled by a first order differential equation. These much more restricted models could in principle be represented by SDCMs. The dynamic structural causal models of (Rubenstein et al., 2018) have been developed to model the asymptotic behavior of an ordinary differential equation under non-constant interventions and assume that the asymptotic behavior does not depend on the initial condition.

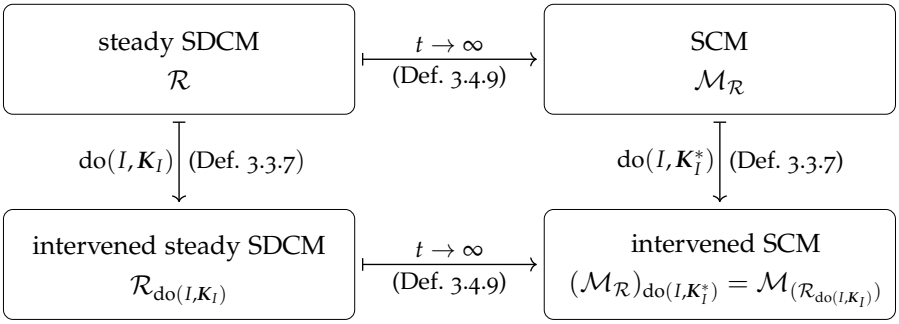


Figure 3.1: This diagram shows that, under certain convergence assumptions, equilibration (left-to-right in the diagram) commutes with intervention (top-to-bottom in the diagram). The precise statement is made explicit in Theorem 3.4.18.

addition, we show that this equilibration operation commutes with intervention (as in Figure 3.1), and naturally maps the graph of the SDCM to the graph of the SCM. This provides a different perspective on what Dash (2005) calls the "violation of the equilibration-manipulation commutability property". Our formalism allows us to generalize the main result of Mooij, Janzing, and Schölkopf (2013), which states that certain causally interpreted systems of ODEs can be equilibrated to SCMs, in several directions: (i) we replace the deterministic setting with a more general stochastic setting, that is, we can deal with randomness in the initial conditions and in the parameters, (ii) we allow the order of the equations of the dynamical model to be arbitrary, including zeroth-order, rather than restricting to first-order differential equations only, and (iii) we drop the strong assumption that the dynamical model needs to have a single static equilibrium that is independent of the initial condition.

By no longer restricting to first-order dynamical systems, we arrive at a more natural causal interpretation of systems of higher-order RDEs, like the coupled harmonic oscillator. Thereby, we circumvent questions like "does position cause velocity, or does velocity cause position, or both?". However, allowing for zeroth-order dynamic structural equations leads to additional technical challenges that are absent when solving first-order RDEs. Indeed, the initial conditions of the solutions may be constrained by the zeroth-order dynamic structural equations, and possibly even by additional "hidden" constraints. We provide sufficient conditions under which the existence and uniqueness of a solution of an SDCM with a given initial condition can be guaranteed. We also provide stronger conditions under which this still holds after certain interventions.

The existence and uniqueness of solutions of an SDCM are of key importance for obtaining Markov properties for SDCMs. By building on a powerful Markov property for SCMs (Forré and Mooij, 2017; Bongers et al., 2021), we derive a Markov property for SDCMs with initial conditions, which enables one to read off (conditional) independencies between the stochastic processes that are solutions of the SDCM, provided the latter are uniquely defined. With a small extension, it can also be applied to the evaluation of the solutions at some specific point in time.

Even if the existence and uniqueness of a solution of an SDCM can be guaranteed, not all solutions of an SDCM equilibrate, in general. For example, a coupled

harmonic oscillator may oscillate indefinitely in the absence of friction. Moreover, the solutions that equilibrate may not always equilibrate to the same equilibrium state. For example, a freely moving particle subject to friction may end up anywhere, depending on its initial position and velocity. In other words, equilibrium states may depend on the initial condition. This is compatible with the recently proposed framework of cyclic SCMs discussed in Chapter 2, which allows for the absence of (or, the presence of multiple) solutions of the structural equations. The intricate connection between the dependence of the equilibrium states of an SDCM on the initial conditions and the solvability properties of the equilibrated SCM sheds new light on the counterintuitive "nonancestral" causal effects in certain "pathological" cyclic SCMs with self-cycles that were first observed by Neal (2000).

The scope of this chapter is limited to establishing the framework of SDCMs and its bridge to SCMs at equilibrium. The importance of this bridge is that, although SDCMs can be used for modeling causal relationships between stochastic processes, inferring such causal models from data may pose certain difficulties. One significant practical drawback of using SDCMs for modeling systems with an unknown dynamics is that obtaining time series data with sufficiently high temporal resolution can be costly, impractical or even impossible.[3] The results of this work enable one to study the causal semantics of the equilibrium states of a large class of random dynamical models in terms of SCMs. In particular, this allows to infer properties of these dynamical models by employing the statistical tools and discovery methods available for static SCMs on equilibrium data.

OUTLINE This chapter is organized as follows: In Section 3.2, we provide the necessary concepts of stochastic processes and random differential equations. In Section 3.3, we introduce the class of structural dynamical causal models, define SCMs as special cases of SDCMs, define interventions, define the graph of an SDCM, discuss initial conditions, study existence and uniqueness of solutions, and derive a Markov property for SDCMs. In Section 3.4, we define the equilibration operation on steady SDCMs, define the graph of the equilibrated SDCM, describe the commutation of the intervention and the equilibration operation, study the inverse problem of finding steady SDCMs with non-trivial dynamics for which all the solutions equilibrate to solutions of the SCM, and discuss subtleties in the causal interpretation of the graph of the equilibrated SDCM. We conclude with a discussion and some open problems in Section 3.5. Proofs are provided in Appendix 3.A.

3.2 PRELIMINARIES

We start off by defining some basic notation and terminology.

3 For example, modern measurement techniques in biology, like RNA sequencing and mass cytometry, enable simultaneous measurements of multiple variables at once in single cells, but at the cost of destroying the cells during the measurement process. This means that it is impossible to obtain time-series measurements for individual cells, although one can take a "snapshot" of the internal states of many single cells at the same point in time.

3.2.1 *Stochastic processes*

In this subsection, we introduce the basic definitions and terminology for stochastic processes (see also Bunke, 1972; Neckel and Rupp, 2013). A *stochastic process* is an $\mathbb{R}^n$-valued function $\boldsymbol{X} : T \times \Omega \to \mathbb{R}^n$, where T is some index set, such that $\boldsymbol{X}_t$ (which denotes $\boldsymbol{X}(t, .)$, also sometimes denoted as $\boldsymbol{X}(t)$) is for each $t \in T$ a random variable[4] on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A random variable $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ can itself be seen as a stochastic process that is constant in time, that is, as the process $\boldsymbol{X} : T \times \Omega \to \mathbb{R}^n$ defined by $\boldsymbol{X}_t(\omega) := \boldsymbol{X}(\omega)$. We always assume that there exists some background probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables and processes are defined. Furthermore, we only consider processes where $T = [t_0, t_1]$ or $T = [t_0, \infty)$ for $t_0 < t_1$ with $t_0, t_1 \in \mathbb{R}$, and the points of T are thought of as representing *time*. For each $\omega \in \Omega$ we have an $\mathbb{R}^n$-valued function $T \to \mathbb{R}^n$ mapping t to $\boldsymbol{X}_t(\omega)$, which is called a *sample path*, or just a *path*, of $\boldsymbol{X}$. We call two stochastic processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ *a.s. equal* to each other, denoted by $\boldsymbol{X} = \boldsymbol{Y}$ a.s., if $\mathbb{P}$-almost surely all sample paths are equal, that is, if there exists a $\mathbb{P}$-null set[5] $N \subseteq \Omega$ such that for all $\omega \in \Omega \setminus N$ and for all $t \in T$ we have $\boldsymbol{X}_t(\omega) = \boldsymbol{Y}_t(\omega)$. We consider stochastic processes, and random variables in particular, only up to a.s. equality.

A family $(\boldsymbol{X}_i)_{i \in \mathcal{I}}$ of stochastic processes for some finite index set $\mathcal{I}$ is called *independent* if for all $k \in \mathbb{N}$ and all k-tuples $(t_1, \ldots, t_k)$ of distinct elements of T the family

$$(\tilde{\boldsymbol{X}}_i)_{i \in \mathcal{I}}$$

of random variables $\tilde{\boldsymbol{X}}_i := ((\boldsymbol{X}_i)_{t_1}, \ldots, (\boldsymbol{X}_i)_{t_k})$ is independent.

We call a stochastic process $\boldsymbol{X}$ *continuous*, if its paths are continuous almost surely, that is, for $\mathbb{P}$-almost every $\omega \in \Omega$ and for all $t \in T$ we have

$$\lim_{s \to t} \boldsymbol{X}_s(\omega) = \boldsymbol{X}_s(\omega)\,.$$

We call a stochastic process $\boldsymbol{X}$ *differentiable*, if its paths are differentiable almost surely, that is, for $\mathbb{P}$-almost every $\omega \in \Omega$ and for all $t \in T$ the derivative

$$\boldsymbol{X}'_t(\omega) := \frac{d\boldsymbol{X}_t}{dt}(\omega) := \lim_{h \to 0} \frac{\boldsymbol{X}_{t+h}(\omega) - \boldsymbol{X}_t(\omega)}{h}$$

exists. The mapping $\boldsymbol{X}' : T \times \Omega \to \mathbb{R}^n$ defines a stochastic process and is called the *derivative* of $\boldsymbol{X}$. Similarly, one can define, if it exists, the n^{th}*-order derivative* of $\boldsymbol{X}$ as the derivative of the $(n-1)^{\text{th}}$-order derivative of $\boldsymbol{X}$, which we also write as $\boldsymbol{X}^{(n)}$, where the zeroth-order derivative of $\boldsymbol{X}$ is $\boldsymbol{X}^{(0)} := \boldsymbol{X}$. We call a stochastic process $\boldsymbol{X}$ *continuously differentiable* or a C^1*-stochastic process*, if its derivative $\boldsymbol{X}'$ exists and is continuous. Similarly, we call $\boldsymbol{X}$ a C^m*-stochastic process*, if its derivatives $\boldsymbol{X}'$, $\boldsymbol{X}''$,

4 Assuming the Borel σ-algebra $\mathfrak{B}(\mathbb{R}^n)$ on $\mathbb{R}^n$, that is, the smallest σ-algebra on $\mathbb{R}^n$ that contains all open n-balls.

5 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A set $N \subseteq \Omega$ is called a $\mathbb{P}$-*null set* if there exists a measurable set $\tilde{N} \in \mathcal{F}$ with $N \subseteq \tilde{N}$ and $\mathbb{P}(\tilde{N}) = 0$.

$\dots, \boldsymbol{X}^{(m)}$ exist and are continuous. In particular, $\boldsymbol{X}$ is a C^0-stochastic process if it is continuous.

Consider a compact interval $T = [t_0, t_1] \subseteq \mathbb{R}$. The space $\mathcal{C}^m(T, \mathbb{R}^n)$ of m times continuously differentiable functions $T \to \mathbb{R}^n$, equipped with the C^m-norm

$$\|\boldsymbol{X}\|^{(m)} := \sum_{k=0}^{m} \sup_{t \in T} \|\boldsymbol{X}^{(k)}(t)\|$$

(where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^n$) is a Polish space, and with its Borel σ-algebra forms a standard measurable space (Kechris, 1995). A C^m-stochastic process $\boldsymbol{X} : T \times \Omega \to \mathbb{R}^n$ can also be seen as a random variable taking values in $\mathcal{C}^m(T, \mathbb{R}^n)$ (Borovkov, 2013). The following functionals (integration, differentiation and evaluation) are continuous, and hence measurable:

$$\begin{aligned}
\iota &: \mathbb{R}^n \times \mathcal{C}^m(T, \mathbb{R}^n) \to \mathcal{C}^{m+1}(T, \mathbb{R}^n) : (\boldsymbol{X}_{[0]}, \boldsymbol{X}) \mapsto \left(t \mapsto \boldsymbol{X}_{[0]} + \int_{t_0}^{t} \boldsymbol{X}(s)\, ds \right) \\
\partial &: \mathcal{C}^{m+1}(T, \mathbb{R}^n) \to \mathcal{C}^m(T, \mathbb{R}^n) : \boldsymbol{X} \mapsto \left(t \mapsto \boldsymbol{X}'(t) \right) \\
\pi &: \mathcal{C}^m(T, \mathbb{R}^n) \to \mathbb{R}^n : \boldsymbol{X} \mapsto \boldsymbol{X}(t_1)\,.
\end{aligned}$$

Furthermore, if we compose a process $\boldsymbol{X} \in \mathcal{C}^n(T, \mathbb{R}^n)$ with a continuous function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^k$, we obtain a process $\boldsymbol{f}(\boldsymbol{X}) \in \mathcal{C}^0(T, \mathbb{R}^k)$.

3.2.2 *Clustered mixed graphs*

In this subsection, we introduce some graphical notions.

A *mixed graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges between the nodes of different types, in our case, $\to, \leftarrow, \leftrightarrow, \dashrightarrow, \dashleftarrow$. If $i \to j$ or $i \dashrightarrow j$ in $\mathcal{G}$, we call i a *parent of* j and denote with $\mathrm{pa}_{\mathcal{G}}(j)$ the set of parents of j (which may include j itself in case $j \dashrightarrow j$ in $\mathcal{G}$). A mixed graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ is a *subgraph* of a mixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $\tilde{\mathcal{V}} \subseteq \mathcal{V}$ and $\tilde{\mathcal{E}} \subseteq \mathcal{E}$.

A *clustered mixed graph* is a triple $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ where $(\mathcal{V}, \mathcal{E})$ is a mixed graph and $\mathcal{P}$ is a partition of the nodes $\mathcal{V}$, such that dashed edges $\dashrightarrow, \dashleftarrow$ only appear between nodes in the same element of $\mathcal{P}$. Each element of $\mathcal{P}$ is called a *cluster* of the clustered mixed graph. A clustered mixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ induces a mixed graph $\mathrm{col}(\mathcal{G})$ with nodes $\mathcal{P}$, a directed edge $K \to L$ for $K \neq L$ iff there is a directed edge $k \to l$ in $\mathcal{G}$ for some $k \in K, l \in L$, and a bidirected edge $K \leftrightarrow L$ for $K \neq L$ iff there is a bidirected edge $k \leftrightarrow l$ in $\mathcal{G}$ for some $k \in K, l \in L$. This construction can be thought of as "collapsing" the clusters in the clustered mixed graph into nodes and subsequently removing self-cycles.

3.2.3 *Random differential equations*

In this subsection, we give a brief overview of some key aspects of random differential equations (for more details, see Bunke, 1972; Neckel and Rupp, 2013). Random

differential equations (RDEs) are similar to ordinary differential equations (ODEs), but can deal with randomness in the initial conditions and in the parameters. Due to their close connection to ODEs they can be analyzed by use of methods that are analogous to those in the theory of ODEs (Bunke, 1972). Their formalism is conceptually easier than the formalism of the white-noise driven stochastic differential equations (SDEs), while still being applicable to those systems via the generalized Doss-Sussmann correspondence (see Jentzen and Kloeden, 2011; Neckel and Rupp, 2013). They have been used for many years in a wide range of applications (see, for example, Bunke, 1972; Soong, 1973; Sobczyk, 1991; Neckel and Rupp, 2013; Han and Kloeden, 2017; Liu et al., 2020).

A stochastic process $\boldsymbol{X} : T \times \Omega \to \mathbb{R}^d$ is a *solution* of a *(first-order) random differential equation*

$$\boldsymbol{X}' = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E}), \tag{3.1}$$

where $\boldsymbol{f} : \mathbb{R}^d \times \mathbb{R}^e \to \mathbb{R}^d$ is a measurable function and $\boldsymbol{E} : T \times \Omega \to \mathbb{R}^e$ a stochastic process, if for $\mathbb{P}$-almost every $\omega \in \Omega$ the (first-order) ordinary differential equation[6]

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

holds for all $t \in T$. An *initial condition* of the RDE (3.1) is a tuple $(t_0, \boldsymbol{X}_{[0]})$ that specifies those solutions $\boldsymbol{X}$ of the RDE (3.1) that satisfy for $\mathbb{P}$-almost every $\omega \in \Omega$

$$\boldsymbol{X}_{t_0}(\omega) = \boldsymbol{X}_{[0]}(\omega)$$

at the initial time t_0. Since every n^{th}-order ODE can be rewritten as a system of first-order ODEs, the general form of the random differential equation (3.1) can be used to express analogously all the *n^{th}-order random differential equations*.[7]

The inclusion of randomness in the equations can be classified into two basic types. The first type consists of *randomness in the initial conditions*, that is, the initial conditions are not a.s. equal to a constant deterministic process. The second type consists of *randomness in the parameters*, that is, the process $\boldsymbol{E}$ is not a.s. equal to a deterministic stochastic process. Of course, a combination of both types can hold. In particular, an RDE together with an initial condition reduces to an *initial value problem* for ODEs if it has no randomness in both the initial conditions and the parameters.

If the stochastic process $\boldsymbol{E}$ is continuous, sufficient conditions that guarantee the existence and uniqueness of solutions for any initial condition can be found in Bunke (1972) and Kloeden and Platen (1992). These results are similar to the uniqueness and existence theorems for ODEs (Coddington and Levinson, 1955).

6 These ordinary differential equations are also called *explicit ordinary differential equations* (Ascher and Petzold, 1998). Similarly, the random differential equations (3.1) are also called *explicit random differential equations*.

7 Furthermore, explicit time-dependence of $\boldsymbol{f}$ can be incorporated by adding a dummy variable with t with dynamics $t' = 1$ and initial condition $t_{[0]} = 0$.

Figure 3.2: Two masses coupled by a spring, freely drifting in space (left, see Example 3.2.1) and with one of the masses attached to a fixed point (right, see Example 3.2.2).

Example 3.2.1 (Two masses coupled by a spring)**.** *Consider a one-dimensional system of two point masses m_1 and m_2 with positions X_1 and X_2 respectively that are coupled by an ideal spring with spring constant $\kappa_1 > 0$ and equilibrium length $L_1 > 0$ under influence of friction with friction coefficients $b_1, b_2 \geq 0$ respectively (see Figure 3.2 (left)). The equations of motion of this system, whose derivation can be found in physics textbooks, are given by the second-order random differential equations*

$$\begin{cases} X_1'' = \dfrac{\kappa_1}{m_1}(X_2 - X_1 - L_1) - \dfrac{b_1}{m_1}X_1' \\ X_2'' = \dfrac{\kappa_1}{m_2}(X_1 - X_2 + L_1) - \dfrac{b_2}{m_2}X_2' \,. \end{cases}$$

Randomness may enter the system via the initial condition

$$(t_0, (X_1(t_0), X_1'(t_0), X_2(t_0), X_2'(t_0)))$$

or via the parameters. For example, instead of assuming that the length L_1 has a fixed value, we can assume that it is an exogenous random variable distributed according to some distribution. The system of equations then forms an RDE.

In this chapter, we propose a modeling class that allows to model the causal semantics of stochastic processes with RDEs in an unambiguous way. The following example illustrates that modeling interventions on RDEs, and thereby grounding their causal semantics, is not a completely trivial matter.

Example 3.2.2 (Two masses coupled by a spring, continued)**.** *Consider again the RDE that describes the two masses coupled by an ideal spring from Example 3.2.1. These equations denote a symmetric relation, that is, for both equations X_1 can be expressed in terms of X_2, and vice versa. The causal relations between the processes X_1 and X_2 are not inherently implied by the form of the equations. For example, what happens to X_2 if we fix the mass m_1 to a fixed wall, say, at $X_1 = 0$ (see Figure 3.2 (right))? The corresponding RDE for X_1 and X_2 is then given by*

$$\begin{cases} X_1 = 0 \\ X_2'' = \dfrac{\kappa_1}{m_2}(X_1 - X_2 + L_1) - \dfrac{b_2}{m_2}X_2' \,. \end{cases}$$

In both cases we implicitly assumed that each mass has its own equation of motion, that is, the first and second equation determine the motion of the mass m_1 and m_2, in terms of the processes X_1 and X_2, respectively. Therefore, the intervention of fixing the mass m_1 to the wall is accomplished by changing only the equation for m_1 to the equation $X_1 = 0$. If

instead we had changed the other equation to $X_1 = 0$, then, as one can easily verify, X_2 would always be fixed, which does not correspond to the expected physical behavior. This additional "structure" of knowing which RDE determines the dynamics of which process is not "intrinsically" defined by the RDE. Moreover, RDEs usually do not include zeroth-order equations (also referred to as "algebraic equations"), such as $X_1 = 0$. Allowing for RDEs of arbitrary order, including zeroth-order, allows to model a wide range of interventions on these models. For example, instead of fixing the mass m_1 to the fixed wall at $X_1 = 0$ we could fix it to a wall that is driven by some external force, such as $X_1 = A\sin(2\pi f t)$ for some $A, f > 0$.

3.3 STRUCTURAL DYNAMICAL CAUSAL MODELS

In this section, we introduce the class of structural dynamical causal models (SDCMs) that allows to formally specify causal semantics for any RDE of arbitrary order (including zeroth-order). We organize the differential equations of the RDEs in a structural way, similar to how this is done for structural causal models, such that each differential equation expresses the causal mechanism that governs the dynamics of a single stochastic process (corresponding to a single component of the system). This allows us to model stochastic idealized interventions targeting certain components in dynamical models, similarly to how this is done for SCMs.

We start in Section 3.3.1 with introducing the notation and terminology that will be used throughout this chapter. In Section 3.3.2, we formally define SDCMs and their solutions. In Section 3.3.3, we formalize the causal semantics of SDCMs in terms of stochastic "perfect" interventions. In Section 3.3.4, we introduce and discuss a graphical representation for SDCMs. In Section 3.3.5, we discuss the initial conditions and how these relate to the existence of solutions. In Section 3.3.6, we state results about the existence and uniqueness of solutions of certain classes of SDCMs. We finish in Section 3.3.7 by deriving a Markov property for SDCMs with initial conditions, suitable for both the solutions of the SDCM and the evaluation of the solutions at any point in time.

3.3.1 *Notation and terminology*

Let $\mathcal{I} = \{1, \dots, d\}$ be a finite index set and $\boldsymbol{\mathcal{X}} = \prod_{i\in\mathcal{I}} \mathcal{X}_i$ the product of the domains of the components of a system, where domain $\mathcal{X}_i = \mathbb{R}^{d_i}$ encodes the range of possible values that the i^{th} component can take. The stochastic process $\boldsymbol{X} = (X_1, \dots, X_d) : T \times \Omega \to \boldsymbol{\mathcal{X}}$ has component processes $X_i : T \times \Omega \to \mathcal{X}_i$.

Let $i \in \mathcal{I}$ and $n_i \in \mathbb{N}_0$. If for the i^{th} component X_i the n_i^{th}-order derivative exists, then the *complete n_i^{th}-order derivative* of X_i, defined as the stochastic process $\overline{X_i}^{(n_i)} := (X_i, X_i', X_i'', \dots, X_i^{(n_i)}) : T \times \Omega \to \mathcal{X}_i^{n_i+1}$, is the tuple of all the derivatives of X_i up to and including order n_i. We adopt a similar notation for the values in $\mathcal{X}_i^{n_i+1}$, that is, $\bar{x}_i^{(n_i)} \in \mathcal{X}_i^{n_i+1}$. Each component $X_i^{(k_i)}$ of $\overline{X_i}^{(n_i)}$, or similarly $x_i^{(k_i)}$ of $\bar{x}_i^{(n_i)}$,

corresponds to an index $i^{(k_i)}$, which gives the index set $\overline{i}^{(n_i)} := \{i^{(k_i)} : 0 \leq k_i \leq n_i\}$ for $\overline{X}_i^{(n_i)}$, where the index $i^{(0)}$ is also written as i.

Let $\boldsymbol{n} = (n_1, \ldots, n_d) \in \mathbb{N}_0^{\mathcal{I}}$ be a tuple. If the n_i^{th}-order derivative of X_i exists for every $i \in \mathcal{I}$, then the *$\boldsymbol{n}^{th}$-order derivative* of $\boldsymbol{X}$ is defined as the stochastic process $\boldsymbol{X}^{(\boldsymbol{n})} := (X_1^{(n_1)}, \ldots, X_d^{(n_d)}) : T \times \Omega \to \boldsymbol{\mathcal{X}}$ and the *complete $\boldsymbol{n}^{th}$-order derivative* of $\boldsymbol{X}$ is defined as the stochastic process $\overline{\boldsymbol{X}}^{(\boldsymbol{n})} := (\overline{X_1}^{(n_1)}, \overline{X_2}^{(n_2)}, \ldots, \overline{X_d}^{(n_d)}) : T \times \Omega \to \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1}$, where $\boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1} := \prod_{i=1}^{d} \mathcal{X}_i^{n_i+1}$. We adopt a similar notation for the values in $\boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1}$, that is, $\bar{\boldsymbol{x}}^{(\boldsymbol{n})} \in \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1}$. Similarly, each component $X_i^{(k_i)}$ of $\overline{\boldsymbol{X}}^{(\boldsymbol{n})}$ corresponds to an index $i^{(k_i)}$ which gives the index set $\overline{\mathcal{I}}^{(\boldsymbol{n})} := \bigcup_{i \in \mathcal{I}} \overline{i}^{(n_i)}$ for $\overline{\boldsymbol{X}}^{(\boldsymbol{n})}$.

For a subset $I := \{i_1, \ldots, i_k\} \subseteq \mathcal{I}$ we will use the notation $\boldsymbol{n}_I := (n_{i_1}, \ldots, n_{i_k})$ and write $\boldsymbol{\mathcal{X}}_I = \prod_{i \in I} \mathcal{X}_i$ and $\boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I+1} = \prod_{i \in I} \mathcal{X}_i^{n_i+1}$. For the I^{th} components of the process $\boldsymbol{X}$ and the complete $\boldsymbol{n}^{\text{th}}$-order derivative $\overline{\boldsymbol{X}}^{(\boldsymbol{n})}$, we write $\boldsymbol{X}_I := (X_{i_1}, \ldots, X_{i_k})$ and $\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I)} := (\overline{X}_{i_1}^{(n_{i_1})}, \ldots, \overline{X}_{i_k}^{(n_{i_k})})$ respectively. Similarly, for the values in $\boldsymbol{\mathcal{X}}_I$ and $\boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I+1}$, we write $\boldsymbol{x}_I := (x_{i_1}, \ldots, x_{i_k}) \in \boldsymbol{\mathcal{X}}_I$ and $\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I)} := (\bar{x}_{i_1}^{(n_{i_1})}, \ldots, \bar{x}_{i_k}^{(n_{i_k})}) \in \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I+1}$ respectively.

In this notation, a stochastic process X is a *C^n-stochastic process*, if its complete n^{th}-order derivative $\overline{X}^{(n)}$ exists and is continuous. Similarly, we call a stochastic process $\boldsymbol{X}$ a *$C^{\boldsymbol{n}}$-stochastic process*, if its complete $\boldsymbol{n}^{\text{th}}$-order derivative $\overline{\boldsymbol{X}}^{(\boldsymbol{n})}$ exists and is continuous. We denote by $C^{\boldsymbol{n}}(T, \boldsymbol{\mathcal{X}})$ the space of $C^{\boldsymbol{n}}$-stochastic processes. For $T = [t_0, t_1] \subseteq \mathbb{R}$ compact, the space $C^{\boldsymbol{n}}(T, \boldsymbol{\mathcal{X}})$ forms a standard measurable space with Borel σ-algebra given by the $C^{\boldsymbol{n}}$-norm

$$\|\boldsymbol{X}\|^{(\boldsymbol{n})} := \sum_{i \in \mathcal{I}} \sum_{k=0}^{n_i} \sup_{t \in T} \|X_i^{(k)}(t)\| .$$

3.3.2 *Structural dynamical causal models and their solutions*

Informally, we think of an SDCM as an SCM where we replace the random variables of the SCM by stochastic processes and their derivatives, and where each structural equation of the SCM becomes a random differential equation of arbitrary order. This generalizes the class of SCMs to the continuous time domain and enables a causal semantics for a broad range of random dynamical models. In this chapter, we closely follow the terminology for SCMs of Chapter 2 and extend it to SDCMs.

Definition 3.3.1 (Structural dynamical causal model)**.** *A* structural dynamical causal model (SDCM) *is a tuple*[8]

$$\mathcal{R} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$$

where

- *$\mathcal{I}$ is a finite index set for endogenous processes,*

8 We often use boldface for variables that have multiple components, that is, which take values in a Cartesian product.

- $\mathcal{J}$ *is a disjoint finite index set for exogenous processes,*
- $\boldsymbol{\mathcal{X}} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$ *is the product of the domains of the endogenous processes, where each domain* $\mathcal{X}_i = \mathbb{R}^{d_i}$,
- $\boldsymbol{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ *is the product of the domains of the exogenous processes, where each domain* $\mathcal{E}_j = \mathbb{R}^{e_j}$,
- $\boldsymbol{n} = (n_i)_{i \in \mathcal{I}} \in \mathbb{N}_0^{\mathcal{I}}$ *is the* order tuple,
- $\boldsymbol{f} : \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ *is a measurable function that specifies the* dynamic causal mechanism,
- $\boldsymbol{E} : T \times \Omega \to \boldsymbol{\mathcal{E}}$ *is an exogenous stochastic process with independent components, that is,* $(E_j)_{j \in \mathcal{J}}$ *is independent.*

The solutions of a structural dynamical causal model in terms of stochastic processes are defined by the associated dynamic structural equations.

Definition 3.3.2 (Solution of an SDCM)**.** *A stochastic process* $\boldsymbol{X} : T \times \Omega \to \boldsymbol{\mathcal{X}}$ *is a* solution *of the* dynamic structural equations *(dynamic SEs) associated to SDCM* $\mathcal{R}$*,*

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

if $\boldsymbol{X}$ *is a* $C^{\boldsymbol{n}}$*-stochastic process, and for* $\mathbb{P}$*-almost every* $\omega \in \Omega$ *the ordinary differential equations*[9]

$$\boldsymbol{X}_t(\omega) = \boldsymbol{f}\big(\overline{\boldsymbol{X}}_t^{(\boldsymbol{n})}(\omega), \boldsymbol{E}_t(\omega)\big)$$

hold for all $t \in T$*.*

The value n_i of the order tuple $\boldsymbol{n}$ denotes the highest-order derivative of X_i that may occur in the dynamic structural equations. Note that taking higher n_i's will in general reduce the set of possible solutions, due to additional imposed smoothness constraints on the solutions. In contrast to the common way of writing RDEs (see equation (3.1)), the (higher-order) derivatives of the endogenous processes of an SDCM always appear on the right-hand side of the dynamic SEs.[10] This notation explicitly allows us to model zeroth-order dynamic structural equations, that is, equations that contain no derivatives of order one or higher, in other words, random algebraic equations.

In particular, if *all* dynamic structural equations are of zeroth order and the exogenous stochastic processes in the model are constant in time (that is, random variables), then the structural dynamical causal model reduces to a structural causal model (see Chapter 2). In contrast to Definition 2.2.1, we define an SCM here in terms of an exogenous random variable instead of an exogenous distribution.

9 These equations are called *implicit ordinary differential equations* if the Jacobian matrix $\frac{\partial f(\bar{x}^{(n)}, e)}{\partial x^{(n)}}$ is nonsingular for all its argument values in an appropriate domain, otherwise they are called *differential-algebraic equations* (Ascher and Petzold, 1998).

10 For every RDE of the form $\boldsymbol{X}' = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$ with $\boldsymbol{f}$ and $\boldsymbol{E}$ continuous, there exists an SDCM with the same solutions: $\boldsymbol{X}$ is a solution of the RDE if and only if it is a solution of the SDCM $\mathcal{R}$ with the dynamic SE $\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}' + \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$, as long as $\boldsymbol{n} = 1$ (since all solutions of the RDE must be continuously differentiable).

Definition 3.3.3 (Structural causal model)**.** *A* structural causal model (SCM) *is a tuple*

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

such that $\langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \mathbf{0}, \boldsymbol{f}, \boldsymbol{E} \rangle$ *is an SDCM with* $\boldsymbol{E}$ *a random variable.*

That is, we can identify SCMs with certain special cases of SDCMs. Similarly, we can identify the solutions of an SCM (see Definition 2.2.3) with the (constant) solutions of the corresponding SDCM. The following definition is equivalent to Definition 3.3.2 when the latter is applied to an SCM.

Definition 3.3.4 (Solution of an SCM)**.** *A random variable* $\boldsymbol{X} : \Omega \to \mathcal{X}$ *is a* solution *of the* structural equations *associated to SCM* $\mathcal{M}$*,*

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

if for $\mathbb{P}$*-almost every* $\omega \in \Omega$

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

holds.

Similar to the structural equations of an SCM (Woodward, 2003; Pearl, 2009), the dynamic structural equations of an SDCM model the underlying causal mechanisms in a structural way, that is, each dynamic structural equation expresses a specific endogenous process (on the left-hand side) in terms of a dynamic causal mechanism depending on certain processes and their derivatives (on the right-hand side). It is this additional structure, which allows us to explicitly model the causal semantics, that distinguishes structural dynamical causal models from dynamical models such as ODEs and RDEs.[11] Allowing for zeroth and higher-order derivatives of X_i in the dynamic structural equations gives rise to a broad range of random dynamical models that can be described by an SDCM, ranging from ODEs (including first-order ODEs as in (Mooij, Janzing, and Schölkopf, 2013)), RDEs (as in Section 3.2.3) and more general random dynamical systems such as partially equilibrated systems (as in (Iwasaki and Simon, 1994)).

Example 3.3.5 (Damped coupled harmonic oscillator)**.** *Consider a one-dimensional system of* d *point masses* $m_i > 0$ $(i = 1, \dots, d)$ *with positions* $X_i \in \mathbb{R}$*, which are coupled by ideal springs, with spring constants* $\kappa_i > 0$ *and equilibrium lengths* $L_i > 0$ $(i = 1, \dots, d-1)$*, under influence of friction with friction coefficients* $b_i \geq 0$ $(i = 1, \dots, d)$ *(see Figure 3.3 left). This system can be modeled by the SDCM*[12]

$$\mathcal{R} = \langle \{1, \dots, d\}, \{1, \dots, d-1\}, \mathbb{R}^d, \mathbb{R}^{d-1}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$$

11 The importance of assigning a differential equation to an endogenous variable was already observed in (Mooij, Janzing, and Schölkopf, 2013).

12 We abuse notation here; more formally, we should use an index set for $\mathcal{J}$ that is disjoint from $\mathcal{I}$, for example, $\{\tilde{1}, \dots, \tilde{d} - 1\}$.

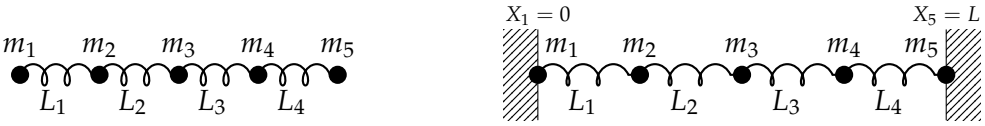


Figure 3.3: Damped coupled harmonic oscillator model $\mathcal{R}$ of Example 3.3.5 (left) and the intervened model $\mathcal{R}_{\mathrm{do}(\{1,5\},(0,L))}$ of Example 3.3.8 (right), both for $d = 5$.

with order tuple $\boldsymbol{n} := (2, \ldots, 2)$*, where the exogenous process* $\boldsymbol{E} = \boldsymbol{L} := (L_1, \ldots, L_{d-1})$ *is constant in time (that is, a random variable), and the causal mechanism is specified by the dynamic structural equations*

$$\begin{cases} X_1 = f_1(\overline{\boldsymbol{X}}^{(\boldsymbol{n})}, \boldsymbol{L}) := X_2 - L_1 - \dfrac{b_1}{\kappa_1} X_1' - \dfrac{m_1}{\kappa_1} X_1'' \\ X_i = f_i(\overline{\boldsymbol{X}}^{(\boldsymbol{n})}, \boldsymbol{L}) := \dfrac{\kappa_i}{\kappa_i + \kappa_{i-1}}(X_{i+1} - L_i) + \dfrac{\kappa_{i-1}}{\kappa_i + \kappa_{i-1}}(X_{i-1} + L_{i-1}) \\ \qquad\qquad - \dfrac{b_i}{\kappa_i + \kappa_{i-1}} X_i' - \dfrac{m_i}{\kappa_i + \kappa_{i-1}} X_i'' \qquad (i = 2, \ldots, d-1) \\ X_d = f_d(\overline{\boldsymbol{X}}^{(\boldsymbol{n})}, \boldsymbol{L}) := X_{d-1} + L_{d-1} - \dfrac{b_d}{\kappa_{d-1}} X_d' - \dfrac{m_d}{\kappa_{d-1}} X_d'' \,. \end{cases}$$

The motion of the masses, in terms of their positions X_i*, velocities* X_i' *and accelerations* X_i''*, is described by a separate equation of motion for each mass. For the case* $d = 2$*, this SDCM* $\mathcal{R}$ *has the same solutions as those described by the RDE in Example 3.2.1.*

The following example motivates why we only consider processes as solutions of SDCMs in case they satisfy the smoothness conditions.

Example 3.3.6 (Sufficient smoothness of the solutions)**.** *Let* $\mathcal{R} = \langle \{1\}, \emptyset, \mathcal{X}, \mathcal{E}, n, f, E \rangle$ *be the SDCM with* $\mathcal{X} = \mathbb{R}$*,* $\mathcal{E}$ *the singleton* $\{*\}$*,* $n = 0$*, the dynamic causal mechanism* $f : \mathcal{X} \times \mathcal{E} \to \mathcal{X}$ *given by* $f(x, e) = x - x^2 + 1$*, and* E *the trivial exogenous process. The zeroth-order dynamic structural equation associated to* $\mathcal{R}$ *reads*

$$X = X - X^2 + 1 \,.$$

This dynamic structural equation does not depend on any exogenous process. The set of endogenous processes $X : T \times \Omega \to \mathbb{R}$ *that satisfy the dynamic structural equation consists of all stochastic processes in* $\{-1, 1\}^{T \times \Omega}$*. Most of the stochastic processes in* $\{-1, 1\}^{T \times \Omega}$ *are not continuous. The solutions of* $\mathcal{R}$ *are exactly those processes in* $\{-1, 1\}^{T \times \Omega}$ *that are* C^0*-stochastic processes. These are the processes that are constant in time, that is, the random variables of* $\{-1, 1\}^{\Omega}$*. In particular, the solutions of the SDCM* $\mathcal{R}$ *correspond exactly to the solutions of the SCM described by the above structural equation.*

3.3.3 *Interventions*

Interventions on a structural dynamical causal model can be modeled in different ways. We consider here a stochastic version of *perfect* interventions[13] on the endogenous processes (Eberhardt and Scheines, 2007) that are analogous to stochastic perfect interventions in structural causal models (Pearl, 2009; Eberhardt, 2014). A stochastic perfect intervention on some endogenous process forces the intervened process to be equal to a given independent exogenous process. More generally, we model a stochastic perfect intervention on a subset $I := \{i_1, \dots, i_k\} \subseteq \mathcal{I}$ of the endogenous processes by forcing those processes $\boldsymbol{X}_I$ to be equal to the intervened processes $\boldsymbol{K}_I$, by changing the model such that the corresponding dynamical structural equations become $\boldsymbol{X}_I = \boldsymbol{K}_I$. The process $\boldsymbol{K}_I$ is treated as an independent exogenous process, such that all its components K_i are mutually independent and independent from all the other exogenous processes that were already present in the model in the absence of the intervention. The dynamic causal mechanisms of the other endogenous processes $\mathcal{I} \setminus I$ are untouched and their dynamics are still specified by the same dynamic structural equations associated to those processes in the absence of the intervention, that is[14]

$$\boldsymbol{X}_{\setminus I} = \boldsymbol{f}_{\setminus I}(\overline{\boldsymbol{X}}^{(n)}, \boldsymbol{E}) \,.$$

This yields the following formal definition of an intervened structural dynamical causal model.

Definition 3.3.7 (Stochastic perfect intervention on an SDCM)**.** *Consider an SDCM* $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, n, \boldsymbol{f}, \boldsymbol{E} \rangle$*, a subset* $I \subseteq \mathcal{I}$*, and a stochastic process* $\boldsymbol{K}_I : T \times \Omega \to \boldsymbol{\mathcal{X}}_I$ *such that* $((K_i)_{i \in I}, (E_j)_{j \in \mathcal{J}})$ *is independent. The* stochastic perfect intervention $\mathrm{do}(I, \boldsymbol{K}_I)$ *maps* $\mathcal{R}$ *to the SDCM*[15]

$$\mathcal{R}_{\mathrm{do}(I, \boldsymbol{K}_I)} := \langle \mathcal{I}, I \cup \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{E}}, n, \tilde{\boldsymbol{f}}, (\boldsymbol{K}_I, \boldsymbol{E}) \rangle \,,$$

where the intervened causal mechanism $\tilde{\boldsymbol{f}} : \boldsymbol{\mathcal{X}}^{n+1} \times (\boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{E}}) \to \boldsymbol{\mathcal{X}}$ *is given by*

$$\tilde{f}_i(\bar{\boldsymbol{x}}^{(n)}, (\boldsymbol{e}_I, \boldsymbol{e}_{\mathcal{J}})) = \begin{cases} f_i(\bar{\boldsymbol{x}}^{(n)}, \boldsymbol{e}_{\mathcal{J}}) & i \in \mathcal{I} \setminus I \\ e_i & i \in I \,. \end{cases} \tag{3.2}$$

We call a stochastic perfect intervention $\mathrm{do}(I, \boldsymbol{K}_I)$ *a* perfect intervention *if* $\boldsymbol{K}_I$ *is a deterministic stochastic process (that is, if it does not depend on* ω*).*

This definition explicitly exposes a hitherto implicit but crucial modeling assumption: exogenous processes are not caused by endogenous processes. Indeed, no

13 These are also referred to as *ideal*, *hard*, *structural*, *surgical*, *atomic* (Eberhardt and Scheines, 2007) or *independent* (Korb et al., 2004) interventions.

14 For $I \subseteq \mathcal{I}$ we adopt the notation $\setminus I$ for $\mathcal{I} \setminus I$.

15 We abuse notation here; more formally, we should make a disjoint copy $\tilde{I} := \{\tilde{i} : i \in I\}$ and use $\tilde{I} \cup \mathcal{J}$ as the new exogenous index set instead of $I \cup \mathcal{J}$, to keep the endogenous indices $\mathcal{I}$ and the exogenous indices $\tilde{I} \cup \mathcal{J}$ disjoint.

stochastic perfect intervention on any subset of the endogenous processes will lead to a change in any of the exogenous processes.

Example 3.3.8. *Consider the damped coupled harmonic oscillator represented by the SDCM $\mathcal{R}$ of Example 3.3.5. Performing the perfect interventions on the masses m_1 and m_d by fixing m_1 and m_d to the walls at $X_1 = 0$ and $X_d = L > 0$, respectively, (see Figure 3.3 (right)) yields the model $\mathcal{R}_{\mathrm{do}(\{1,d\},(0,L))}$ with the dynamic structural equations*

$$\begin{cases} X_1 = 0 \\ X_i = \dfrac{\kappa_i}{\kappa_i + \kappa_{i-1}}(X_{i+1} - L_i) + \dfrac{\kappa_{i-1}}{\kappa_i + \kappa_{i-1}}(X_{i-1} + L_{i-1}) \\ \qquad\quad - \dfrac{b_i}{\kappa_i + \kappa_{i-1}} X_i' - \dfrac{m_i}{\kappa_i + \kappa_{i-1}} X_i'' \qquad (i = 2, \dots, d-1) \\ X_d = L\,. \end{cases}$$

It is clear from the definition that performing stochastic perfect interventions on disjoint subsets of the endogenous processes commutes. In case of overlap, the dynamic structural equations of the overlapping intervention targets are determined by the most recent intervention applied to them.

As a special case, Definition 3.3.7 reduces to the usual notion of (stochastic) perfect intervention on SCMs (see also Definition 2.2.12).

Definition 3.3.9 (Stochastic perfect intervention on an SCM)**.** *Consider an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \boldsymbol{E} \rangle$, a subset $I \subseteq \mathcal{I}$, and a random variable $\boldsymbol{K}_I : \Omega \to \boldsymbol{\mathcal{X}}_I$ such that $((K_i)_{i\in I}, (E_j)_{j\in\mathcal{J}})$ is independent. The* stochastic perfect intervention $\mathrm{do}(I, \boldsymbol{K}_I)$ *maps $\mathcal{M}$ to the SCM*

$$\mathcal{M}_{\mathrm{do}(I,\boldsymbol{K}_I)} := \langle \mathcal{I}, I \cup \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}, (\boldsymbol{K}_I, \boldsymbol{E}) \rangle\,,$$

where $\tilde{\boldsymbol{f}}$ is defined by equation (3.2)*.*

This provides SDCMs with a causal semantics that is analogous to that of SCMs. The following example illustrates how this resolves the ambiguity of the causal interpretation of the RDE of Example 3.2.2.

Example 3.3.10 (Ambiguous causal interpretation of RDEs)**.** *Consider the SDCM $\mathcal{R}$ of Example 3.3.5 for $d = 2$, with dynamic structural equations given by*

$$\begin{cases} X_1 = X_2 - L_1 - \dfrac{b_1}{\kappa_1} X_1' - \dfrac{m_1}{\kappa_1} X_1'' \\ X_2 = X_1 + L_1 - \dfrac{b_2}{\kappa_1} X_2' - \dfrac{m_2}{\kappa_1} X_2''\,. \end{cases}$$

The solutions of $\mathcal{R}$ correspond exactly to the solutions of the RDE that describes the two masses attached to a spring in Example 3.2.1. Fixing the mass m_1 to the left wall at $X_1 = 0$ (see Figure 3.2 (right)) by performing the stochastic perfect intervention[16] $\mathrm{do}(1, K_1)$ *with*

16 For convenience, we write $\mathrm{do}(i, K_i)$ for a stochastic perfect intervention $\mathrm{do}(I, \boldsymbol{K}_I)$ whenever $I = \{i\}$ for some $i \in \mathcal{I}$.

$K_1 = 0$ on $\mathcal{R}$ gives the intervened model $\mathcal{R}_{\mathrm{do}(1,0)}$ with dynamic structural equations given by

$$\begin{cases} X_1 = 0 \\ X_2 = X_1 + L_1 - \dfrac{b_2}{\kappa_1} X_2' - \dfrac{m_2}{\kappa_1} X_2'' \,. \end{cases}$$

The intervened model $\mathcal{R}_{\mathrm{do}(1,0)}$ has exactly the same solutions as the RDE in Example 3.2.2.

Consider now the SDCM $\tilde{\mathcal{R}}$ that is the same as $\mathcal{R}$ except for its dynamic causal mechanism $\tilde{\boldsymbol{f}}$, for which the associated dynamic structural equations are given by

$$\begin{cases} X_1 = X_2 - L_1 + \dfrac{b_2}{\kappa_1} X_2' + \dfrac{m_2}{\kappa_1} X_2'' \\ X_2 = X_1 + L_1 + \dfrac{b_1}{\kappa_1} X_1' + \dfrac{m_1}{\kappa_1} X_1'' \,. \end{cases}$$

Both models $\mathcal{R}$ and $\tilde{\mathcal{R}}$ have the same solutions as those described by the RDEs in Example 3.2.1. However, the intervened models $\mathcal{R}_{\mathrm{do}(1,0)}$ and $\tilde{\mathcal{R}}_{\mathrm{do}(1,0)}$ have different solutions. Only the model $\mathcal{R}_{\mathrm{do}(1,0)}$ describes the expected physical behavior (see also Example 3.2.2).

Stochastic perfect interventions are only defined for the endogenous processes, but not for their higher-order derivatives. The higher-order derivative processes in an SDCM are always obtained by differentiation of the underlying endogenous processes and hence it suffices to define the stochastic perfect interventions only for those underlying endogenous processes. Allowing for stochastic perfect intervention on both the endogenous processes and some of their higher-order derivatives will generally lead to nonsensible causal behavior, as is illustrated in the following example.

Example 3.3.11 (Modeling higher-order derivatives as separate endogenous processes)**.** *Suppose we model the velocities X_i' of the positions X_i of the masses between the walls in the damped coupled harmonic oscillator of Example 3.3.8 explicitly as separate endogenous processes $V_{i'}$. We could attempt to model this with an SDCM $\tilde{\mathcal{R}}$ for which the dynamic structural equations are given by $X_1 = 0$, $X_d = L$ and*

$$\begin{cases} X_i = \dfrac{\kappa_i}{\kappa_i + \kappa_{i-1}}(X_{i+1} - L_i) + \dfrac{\kappa_{i-1}}{\kappa_i + \kappa_{i-1}}(X_{i-1} + L_{i-1}) - \dfrac{b_i}{\kappa_i + \kappa_{i-1}} V_{i'} - \dfrac{m_i}{\kappa_i + \kappa_{i-1}} V_{i'}' \\ V_{i'} = X_i' \end{cases}$$

for $i = 2, \ldots, d-1$. Performing a stochastic perfect intervention on both the position X_i and the velocity $V_{i'}$ of one of the masses between the walls ($i \in \{2, \ldots, d-1\}$) can lead to unphysical behavior. For example, the perfect intervention $\mathrm{do}(\{2, 2'\}, (0, 1))$ gives an intervened SDCM with a solution that is physically impossible if we keep interpreting X_i as the position and $V_{i'}$ the velocity of the i^{th} mass.

This observation constitutes strong motivation for considering the higher-order derivatives $X_i^{(k_i)}$ (up to and including order n_i) to be aspects of the endogenous process X_i rather than as "causally independent" processes. Thereby, we circumvent modeling velocity as the (instantaneous) cause of position (as in Iwasaki and Simon,

1994), or the other way around. The resulting modeling framework appears more natural than that of (Mooij, Janzing, and Schölkopf, 2013), which is explicitly limited to first-order dynamics and cannot accommodate systems like the damped harmonic oscillator as easily as SDCMs can, as it has to impose restrictions on the possible interventions to deal with this problem.

The higher-order derivatives $X_i^{(k_i)}$ do not always exist for a process X_i. For example, if we force the mass m_1 to follow a Brownian motion[17] K_1 in the spring model $\mathcal{R}$ of Example 3.3.10, then the intervened model $\mathcal{R}_{\mathrm{do}(1,K_1)}$ does not yield a solution (because X_1'' needs to exist and be continuous, which is not the case for $X_1 = K_1$). In practice, we therefore only consider stochastic perfect interventions $\mathrm{do}(I, \boldsymbol{K}_I)$ for which $\boldsymbol{K}_I$ is a $C^{\boldsymbol{n}_I}$-stochastic process.

3.3.4 *Graphs*

We will now define a graphical representation of the structural properties of SDCMs that is inspired by the graphical representation of SCMs (see Section 2.2.2). Where the graph of an SCM describes the functional relationships between the random variables encoded by the structural equations, the graph of an SDCM expresses the functional dependencies between the stochastic processes encoded by the dynamic structural equations.

Typically, for $i \in \mathcal{I}$, the component f_i of the dynamic causal mechanism $\boldsymbol{f}$ only depends on a subset of the (derivatives of the) endogenous and exogenous processes that we call the *functional parents of i*.

Definition 3.3.12 (Functional and integrated parents)**.** *Let $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ be an SDCM. For $k \in \overline{\mathcal{I}}^{(\boldsymbol{n})} \cup \mathcal{J}$ and $i \in \overline{\mathcal{I}}^{(\boldsymbol{n})}$, we call*

1. *k a* functional parent of *i if and only if $i \in \mathcal{I}$ and there does not exist a measurable function[18] $\tilde{f}_i : (\boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1})_{\setminus k} \times \boldsymbol{\mathcal{E}}_{\setminus k} \to \mathcal{X}_i$ such that for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\bar{\boldsymbol{x}}^{(\boldsymbol{n})} \in \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1}$*
$$x_i = f_i(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) \quad \iff \quad x_i = \tilde{f}_i((\bar{\boldsymbol{x}}^{(\boldsymbol{n})})_{\setminus k}, \boldsymbol{e}_{\setminus k})\,;$$
2. *k an* integrated parent of *i if and only if there exists an $\ell \in \mathcal{I}$ such that $k = \ell^{(m_\ell - 1)}$ and $i = \ell^{(m_\ell)}$ for some $0 < m_\ell \leq n_\ell$.*

Exogenous processes have no functional and integrated parents by definition. The integrated parents denote the differential relationships that are satisfied by the endogenous processes. That is, for every $\ell \in \mathcal{I}$ and $0 < m_\ell \leq n_\ell$ we have that $\ell^{(m_\ell - 1)}$ is an integrated parent of $\ell^{(m_\ell)}$, which represents the differential relationship
$$X_\ell^{(m_\ell)} = \frac{d}{dt} X_\ell^{(m_\ell - 1)}\,.$$

17 A stochastic process B on $T = [0, \infty)$ is called a *Brownian motion* if: (i) $B_0 = 0$; (ii) B has independent, stationary increments; (iii) $B_t \sim \mathcal{N}(0, t)$ for all $t > 0$; (iv) B is continuous. In particular, B is not differentiable (see, for example, Theorem 21.17 in Klenke, 2014).

18 For $\boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1} = \prod_{i^{(k_i)} \in \overline{\mathcal{I}}^{(\boldsymbol{n})}} \mathcal{X}_i$, some subset $I \subseteq \overline{\mathcal{I}}^{(\boldsymbol{n})}$ and $k \in \overline{\mathcal{I}}^{(\boldsymbol{n})}$, we denote $(\boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1})_{\setminus I} = \prod_{i^{(k_i)} \in \overline{\mathcal{I}}^{(\boldsymbol{n})} \setminus I} \mathcal{X}_i$ and $(\boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1})_{\setminus k} = \prod_{i^{(k_i)} \in \overline{\mathcal{I}}^{(\boldsymbol{n})} \setminus \{k\}} \mathcal{X}_i$, and similarly for their elements.

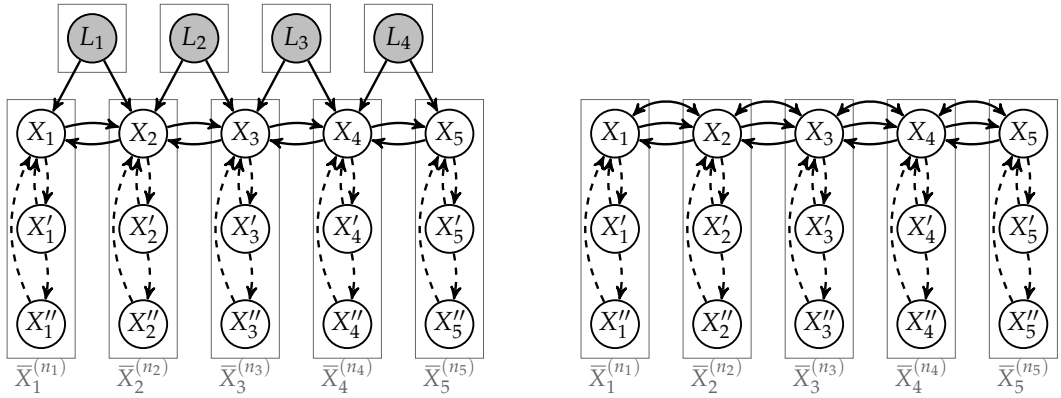


Figure 3.4: Augmented graph (left) and graph (right) of the damped coupled harmonic oscillator model $\mathcal{R}$ of Example 3.3.5 for $d = 5$.

These differential relationships are absent for SCMs, because the endogenous variables are considered static. In contrast to (Iwasaki and Simon, 1994), we express the differential relationships between the endogenous processes by the derivative operator, instead of the integration operator. In general, the integration operator of (Iwasaki and Simon, 1994) is not uniquely defined, since for a particular process there may exist several integrated processes differing by a (possibly random) integration constant. The derivative of a process, however, is always a.s. uniquely defined, if it exists. Hence, for a solution $\boldsymbol{X}$ of an SDCM we can always derive the higher-order derivatives of X_i up to order n_i by repeatedly applying the derivative operator. In this way, we can consider the complete n_i^{th}-order derivative $\overline{X}_i^{(n_i)}$ to encode aspects of the same endogenous process X_i.

The different parental relations can be expressed in a clustered mixed graph, where each cluster represents a complete n_i^{th}-order derivative.

Definition 3.3.13 (Graph and augmented graph)**.** *Let $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ be an SDCM with order tuple $\boldsymbol{n}$. We define:*

1. *the* augmented graph $\mathcal{G}^a(\mathcal{R})$ *of $\mathcal{R}$ as the clustered mixed graph with nodes $\overline{\mathcal{I}}^{(\boldsymbol{n})} \cup \mathcal{J}$ partitioned into* clusters $\overline{i}^{(n_i)} = \{i^{(k_i)} : 0 \leq k_i \leq n_i\}$ *for $i \in \mathcal{I}$ and clusters $\{j\}$ for $j \in \mathcal{J}$, directed edges $k \to l$ if and only if k is functional parent of l in a different cluster, dashed directed edges $k \dashrightarrow l$ if and only if k is a functional or integrated parent of l in the same cluster;*

2. *the* graph $\mathcal{G}(\mathcal{R})$ *of $\mathcal{R}$ as the clustered mixed graph with nodes $\overline{\mathcal{I}}^{(\boldsymbol{n})}$ partitioned into* clusters $\overline{i}^{(n_i)} = \{i^{(k_i)} : 0 \leq k_i \leq n_i\}$ *for $i \in \mathcal{I}$, directed edges $k \to l$ if and only if k is functional parent of l in a different cluster, dashed directed edges $k \dashrightarrow l$ if and only if k is a functional or integrated parent of l in the same cluster, and bidirected edges $k \leftrightarrow l$ if and only if there exists a $j \in \mathcal{J}$ that is a functional parent of both k and l.*

The augmented graph differs from the graph in that it gives an explicit representation of the exogenous processes rather than an implicit one using bidirected

edges. The augmented graph contains no directed edge pointing towards an exogenous process node. The clusters $\overline{i}^{(n_i)} \in \overline{\mathcal{I}}^{(\boldsymbol{n})}$ for $i \in \mathcal{I}$ and $\{j\}$ for $j \in \mathcal{J}$ of the (augmented) graph of an SDCM refer to the complete n_i^{th}-order derivative $\overline{X}_i^{(n_i)}$ and E_j respectively, and are represented by a box around the nodes of the cluster. The graph and augmented graph are illustrated[19] in Figure 3.4 for the damped coupled harmonic oscillator model of Example 3.3.5, where the white and gray nodes represent the endogenous and exogenous processes, respectively. Between the nodes of different clusters there are only functional parental relations. Within a cluster, the higher-order derivatives $i^{(k_i)}$ for $k_i > 0$ of the endogenous processes $i \in \mathcal{I}$ have no functional parents, but have only integrated parents. However, any node $i^{(k_i)}$ with $k_i > 0$ may be a functional parent of another node $j \in \mathcal{I}$; see, for example, the graph of the SDCM $\tilde{\mathcal{R}}$ in Example 3.3.10.[20]

In particular, this definition of the (augmented) graph of an SDCM reduces to the usual notion of the (augmented) graph of an SCM if we ignore the clusters. Indeed, the graph $\mathcal{G}(\mathcal{M})$ of an SCM $\mathcal{M} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \boldsymbol{E} \rangle$ is a mixed graph with nodes $\mathcal{I}$, directed edges $i \to j$ if and only if i is a functional parent of j with $i \neq j$, dashed directed edge $i \dashrightarrow i$ if and only if i is a functional parent of itself, and bidirected edges $i \leftrightarrow j$ if and only if there exists a $k \in \mathcal{J}$ that is a functional parent of both i and j, where we apply Definition 3.3.12 of a functional parent to $\mathcal{M}$ (note that by definition, an SCM has no integrated parents). The augmented graph $\mathcal{G}^a(\mathcal{M})$ of an SCM $\mathcal{M}$ is defined analogously, but the bidirected edges are replaced by exogenous nodes in $\mathcal{J}$ with outgoing directed edges to their functional children.

On the graphs of an SDCM, the operation of a stochastic perfect intervention acts in a simple way.

Proposition 3.3.14 (Graphs of the intervened SDCM)**.** *Let $\mathcal{R}$ be an SDCM and $\mathrm{do}(I, \boldsymbol{K}_I)$ a stochastic perfect intervention for $I \subseteq \mathcal{I}$ a subset and $\boldsymbol{K}_I$ an independent stochastic process. The graph $\mathcal{G}(\mathcal{R}_{\mathrm{do}(I,\boldsymbol{K}_I)})$ of the intervened SDCM $\mathcal{R}_{\mathrm{do}(I,\boldsymbol{K}_I)}$ is the graph $\mathcal{G}(\mathcal{R})$, but without the edges that have an arrowhead pointing towards a node in the intervention target set I. A similar statement holds for the augmented graph $\mathcal{G}^a(\mathcal{R}_{\mathrm{do}(I,\boldsymbol{K}_I)})$.*

The graph and augmented graph of the damped coupled harmonic oscillator model of Example 3.3.8, where we performed the perfect intervention of fixing the endpoint masses to the walls, are illustrated in Figure 3.5. Performing a stochastic perfect intervention on an endogenous process removes all the (bi-)directed edges that point towards the intervened process, including the dashed directed edges within the cluster. The dashed directed edges within the cluster that correspond to the integrated parents, that is, those pointing to a higher-order derivative, indicate that the higher-order derivatives of the intervened endogenous process need to exist for any solution of the model. Hence, we view a stochastic perfect intervention

19 For visualizing the graphs we stick to the common convention of using stochastic processes and random variables with the index as a subscript, instead of using the indices themselves (even when no solutions are defined).

20 A more realistic example could be Faraday's law of induction. In terms of individual point charges: a moving point charge generates a magnetic field, which exerts a force on some other point charge that is proportional to the velocity of the moving point charge.

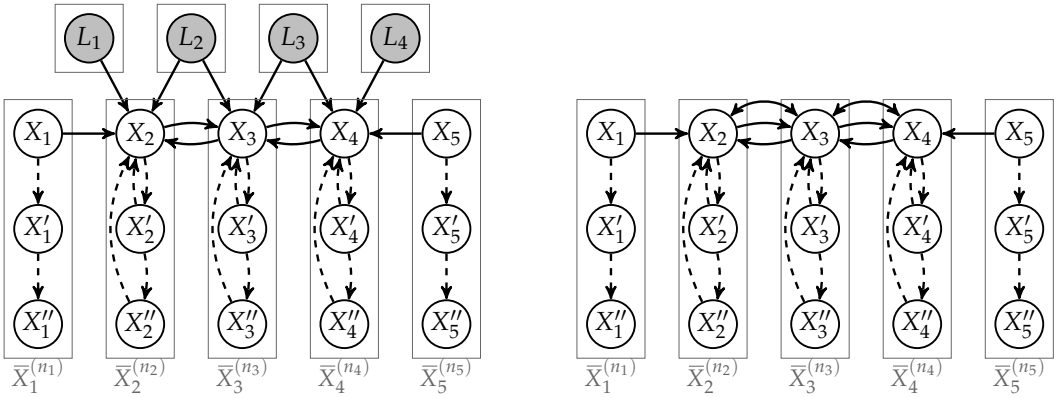


Figure 3.5: Augmented graph (left) and graph (right) of the intervened damped coupled harmonic oscillator model $\mathcal{R}_{\mathrm{do}(\{1,5\},(0,L))}$ of Example 3.3.8 for $d = 5$.

on an endogenous process as an intervention on the whole cluster of the intervened process. We say that there is a *directed edge from cluster I to cluster J* if there exists a directed edge from some $i \in I$ to some $j \in J$. Since a stochastic perfect intervention can be seen as an intervention on the entire associated cluster, the directed edges between the clusters express the direct causal relationships between the clusters. We call a dashed directed edge $i \dashrightarrow i$ in the graph of an SDCM (that is, where i is a functional parent of itself) a *self-cycle at i*. An example of a model where a self-cycle arises is the well-known market equilibrium model from economics, which has been thoroughly discussed in the literature (see, for example, Richardson and Robins, 2014).

Example 3.3.15 (Price, supply and demand)**.** *Let X_P denote the price, X_S denote the supply and X_D the demand of a quantity of a product. The following dynamic structural equations specify an SDCM $\mathcal{R}$ that describes how the demanded and supplied quantities are determined by the price, and how price adjustments occur in the market:*

$$\begin{cases} X_P = X_P + \lambda(X_D - X_S) - X'_P \\ X_S = \beta_S X_P + E_S \\ X_D = \beta_D X_P + E_D \,, \end{cases}$$

where $\boldsymbol{n} := (n_P, n_S, n_D) = (1, 0, 0)$. Here, E_S and E_D are the exogenous influences on the supply and demand respectively, $\beta_S > 0$ is the reciprocal of the slope of the supply curve, $\beta_D < 0$ is the reciprocal of the slope of the demand curve, and $\lambda > 0$ models how fast the price adjusts to market conditions. The graph of this model is depicted in Figure 3.6 (left) and contains a self-cycle at P.

We already encountered several instances of *linear* SDCMs (for example, in Examples 3.3.5 and 3.3.15).

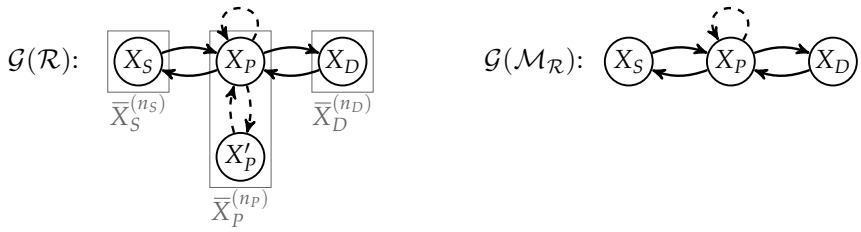


Figure 3.6: Graphs of the price, supply and demand model $\mathcal{R}$ (left) of Example 3.3.15 and the corresponding equilibrated model $\mathcal{M}_\mathcal{R}$ (right) of Example 3.4.17.

Definition 3.3.16 (Linear SDCM)**.** *We call an SDCM $\mathcal{R}$* linear*, if the dynamic causal mechanism $\boldsymbol{f} : \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ is of the form*

$$\boldsymbol{f}(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := B\bar{\boldsymbol{x}}^{(\boldsymbol{n})} + \Gamma \boldsymbol{e}\,,$$

where $B \in \mathbb{R}^{\mathcal{I} \times \overline{\mathcal{I}}^{(\boldsymbol{n})}}$ and $\Gamma \in \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ are matrices.

For a linear SDCM $\mathcal{R}$, a nonzero coefficient B_{ik} for $i, k \in \overline{\mathcal{I}}^{(\boldsymbol{n})}$ such that $i \neq k$ corresponds to a directed edge $k \to i$ in the graph $\mathcal{G}(\mathcal{R})$ (and augmented graph $\mathcal{G}^a(\mathcal{R})$) if i lies in a different cluster than k, and a dashed directed edge $k \dashrightarrow i$ if i lies in the same cluster as k. A coefficient $B_{ii} = 1$ for $i \in \mathcal{I}$ corresponds to a self-cycle $i \dashrightarrow i$. There is a bidirected edge $i \leftrightarrow k$ in the graph $\mathcal{G}(\mathcal{R})$ for $i, k \in \mathcal{I}$ with $i \neq k$ if and only if there exists a $j \in \mathcal{J}$ for which $\Gamma_{ij} \neq 0$ and $\Gamma_{kj} \neq 0$. In the augmented graph $\mathcal{G}^a(\mathcal{R})$, there is a directed edge $j \to i$ for $i \in \mathcal{I}$, $j \in \mathcal{J}$ if and only if $\Gamma_{ij} \neq 0$.

3.3.5 *Initial conditions*

In contrast to RDEs, SDCMs allow for both zeroth and higher-order differential equations. For this reason, the dynamic SEs of SDCMs admit problems that can be quite different from those of RDEs. For example, the order of the initial conditions for SDCMs does not directly relate to the order of the SDCM.

Definition 3.3.17 (Initial condition)**.** *Let $\mathcal{R}$ be an SDCM, $I \subseteq \mathcal{I}$ a subset of the endogenous variables, $\boldsymbol{m}_I = (m_i)_{i \in I} \in \mathbb{N}_0^I$ an order tuple, $t_0 \in T$ and $\overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{m}_I)}$ a random variable taking values in $\boldsymbol{\mathcal{X}}_I^{\boldsymbol{m}_I+1}$. We say that a solution $\boldsymbol{X}$ of $\mathcal{R}$ has* initial condition *$(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{m}_I)})$ if $\overline{\boldsymbol{X}}_I^{(\boldsymbol{m}_I)}(t_0)$ exists and satisfies*

$$\overline{\boldsymbol{X}}_I^{(\boldsymbol{m}_I)}(t_0) = \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{m}_I)}$$

almost surely. Here, $\boldsymbol{m}_I$ is called the order *of the initial condition; for $I = \mathcal{I}$ we also refer to the initial condition as a* full initial condition*, and for $I \subsetneq \mathcal{I}$ as a* partial initial condition*. A solution $\boldsymbol{X}$ of $\mathcal{R}$ with initial condition $(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{m}_I)})$ is called* almost surely unique *if for every solution $\boldsymbol{Y}$ of $\mathcal{R}$ with initial condition $(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{m}_I)})$ we have $\boldsymbol{X} = \boldsymbol{Y}$ a.s..*

For an SDCM for which the dynamic SEs can be rewritten into the form of a system of n_i^{th}-order RDEs (with all $n_i \geq 1$), the full initial conditions of order $\boldsymbol{n} - 1$

of the SDCM correspond exactly with the usually considered initial conditions of this system of RDEs. For example, the solutions of the damped coupled harmonic oscillator of Example 3.3.5 can be a.s. uniquely determined by the full initial conditions of order $\boldsymbol{n} - 1$ (see also Corollary 3.3.28). In general, however, the solutions of an SDCM may not be a.s. uniquely determined by the full initial conditions of order $\boldsymbol{n} - 1$, as the following example illustrates.

Example 3.3.18 (The order of the SDCM and of the initial conditions)**.** *Let $\mathcal{R} = \langle \{1\}, \emptyset, \mathcal{X}, \mathcal{E}, n, f, E \rangle$ be the SDCM with $\mathcal{X} = \mathbb{R}$, $\mathcal{E} = \{*\}$, $n = 1$, the dynamic causal mechanism $f : \mathcal{X}^2 \times \mathcal{E} \to \mathcal{X}$ given by $f(\bar{x}^{(1)}, e) = x - x^2 + (x')^2$, and E the trivial exogenous process. The dynamic structural equation associated to $\mathcal{R}$ reads*

$$X = X - X^2 + (X')^2 \,.$$

This dynamic SE cannot be written as a (first-order) RDE of the form (3.1)*, since it cannot be a.s. uniquely solved for X'. "Solving for" X' leads to two RDEs that are of the form* (3.1)*, namely*

$$X' = X \quad \text{or} \quad X' = -X \,.$$

The solutions of these RDEs are given by $X_t = X_{[0]} e^t$ and $X_t = X_{[0]} e^{-t}$ respectively, where $(0, X_{[0]})$ denotes the initial condition for both RDEs. These processes are also solutions of the SDCM, and one can show that all (continuously differentiable) solutions of $\mathcal{R}$ are of this form. Note that, in principle, we could well have taken the order n arbitrarily high without restricting the set of solutions, because the solutions are C^∞-stochastic processes. If we consider the solutions of $\mathcal{R}$ with an initial condition $(0, \overline{X}^{(0)}_{[0]})$ of order 0, then there are always two solutions with this initial condition that are not a.s. equal to each other, unless $X^{(0)}_{[0]} = 0$. For the initial condition $(0, \overline{X}^{(1)}_{[0]})$ of order 1, we can specify the solution X a.s. uniquely, if it exists. Take for example $\overline{X}^{(1)}_{[0]} = (X_{[0]}, X_{[0]})$, then the solution X with this initial condition is a.s. uniquely given by $X_t = X_{[0]} e^t$. However, an arbitrary initial condition $(0, \overline{X}^{(m)}_{[0]})$ of order m greater or equal to 1 may well be inconsistent with the dynamic structural equations. For example, the initial condition $\overline{X}^{(1)}_{[0]} = (X_{[0]}, 2X_{[0]})$ will not have a solution for $X_{[0]} \neq 0$, since the initial condition $\overline{X}^{(1)}_{[0]} := (X^{(0)}_{[0]}, X^{(1)}_{[0]})$ does not satisfy $(X^{(0)}_{[0]})^2 = (X^{(1)}_{[0]})^2$.

This example illustrates that an arbitrary imposed initial condition may well be inconsistent with the dynamic structural equations.

Definition 3.3.19 (Consistent initial condition)**.** *Let $\mathcal{R}$ be an SDCM and $\boldsymbol{m}_I = (m_i)_{i \in I} \in \mathbb{N}_0^I$ an order tuple for $I \subseteq \mathcal{I}$. We call an initial condition $(t_0, \overline{\boldsymbol{X}}^{(\boldsymbol{m}_I)}_{I,[0]})$ for $\mathcal{R}$* consistent *if there exists a solution of $\mathcal{R}$ with this initial condition.*

In other words, for an initial condition there only exists a solution if and only if the initial condition is consistent. In particular, zeroth-order dynamic structural equations may constrain the initial conditions (of any order) for which a solution exists.

Example 3.3.20 (Zeroth-order dynamic structural equation constraint)**.** *Consider the price, supply and demand model $\mathcal{R}$ of Example 3.3.15 that has order tuple $\boldsymbol{n} = (n_P, n_S, n_D) = (1,0,0)$. The zeroth-order dynamic structural equations of $\mathcal{R}$ are those associated with the supply X_S and the demand X_D processes. Since the solutions of $\mathcal{R}$ satisfy these zeroth-order dynamic structural equations almost surely at every point in time, the consistent full initial conditions $(t_0, \overline{\boldsymbol{X}}_{[0]}^{(m)})$ also need to satisfy the zeroth-order dynamic structural equations almost surely, that is, $X_{[0],S} = \beta_S X_{[0],P} + (E_S)_{t_0}$ and $X_{[0],D} = \beta_D X_{[0],P} + (E_D)_{t_0}$ almost surely.*

By definition, the consistent full initial conditions always need to satisfy the zeroth-order dynamic structural equations of the SDCM. Initial conditions of an order greater than or equal to the order of the SDCM need to satisfy the corresponding dynamic structural equations of the SDCM, as we already saw in Example 3.3.18. Additionally, in general, SDCMs that have higher-order dynamic structural equations may contain *hidden* constraints[21] as the following example illustrates.

Example 3.3.21 (Hidden constraint)**.** *Let $\mathcal{R} = \langle \{1,2\}, \{3\}, \mathbb{R}^2, \mathbb{R}, \boldsymbol{n}, \boldsymbol{f}, E \rangle$ be the SDCM with $\boldsymbol{n} = (0,1)$, the dynamic causal mechanism $\boldsymbol{f} : \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1} \times \mathcal{E} \to \boldsymbol{\mathcal{X}}$ given by $f_1(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, e) := x_2'$ and $f_2(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, e) := e$, and $E := E_3$ some exogenous process. The dynamic structural equations associated to $\mathcal{R}$ read*

$$\begin{cases} X_1 = X_2' \\ X_2 = E\,. \end{cases}$$

This model cannot be written as an RDE,[22] since the Jacobian matrix

$$\frac{\partial \boldsymbol{f}(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e})}{\partial \boldsymbol{x}^{(\boldsymbol{n})}} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2'} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2'} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

is singular everywhere. In order to solve the dynamic SEs we can differentiate the second equation with respect to time to get

$$X_1 = X_2' = E'\,.$$

This SDCM only has solutions if the derivative E' exists. If it exists, then the solutions are given by $X_1 = E'$ and $X_2 = E$. Thus, the solutions satisfy not only the obvious constraint $X_2 = E$, but also need to satisfy the "hidden" constraint $X_1 = E'$. That a solution of the model depends on a derivative of the exogenous variable E cannot happen in a system of RDEs. These constraints imply that every consistent full initial condition $(t_0, \overline{\boldsymbol{X}}_{[0]}^{(\boldsymbol{m})})$ of $\mathcal{R}$ needs to satisfy $X_{[0],1} = E'_{t_0}$ and $X_{[0],2} = E_{t_0}$ almost surely.

21 We refer the reader to the literature on differential-algebraic equations for more details on this, for example, (Ascher and Petzold, 1998).

22 Observe that a higher-order RDE is of the form $\boldsymbol{X}^{(\boldsymbol{n})} = \boldsymbol{g}(\overline{\boldsymbol{X}}^{(\boldsymbol{n}-1)}, \boldsymbol{E})$ for some measurable function $\boldsymbol{g} : \boldsymbol{\mathcal{X}}^{\boldsymbol{n}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ and stochastic process $\boldsymbol{E} : T \times \Omega \to \boldsymbol{\mathcal{E}}$.

After performing a stochastic perfect intervention $\mathrm{do}(I, \boldsymbol{K}_I)$ on an SDCM $\mathcal{R}$, all consistent full initial conditions $(t_0, \overline{\boldsymbol{X}}^{(\boldsymbol{m})}_{[0]})$ must satisfy $\overline{\boldsymbol{X}}^{(\boldsymbol{m}_I)}_{[0],I} = \overline{\boldsymbol{K}}^{(\boldsymbol{m}_I)}_{I}$ almost surely. For example, the consistent initial conditions $(t_0, (\overline{X}^{(1)}_{[0],0}, \ldots, \overline{X}^{(1)}_{[0],d}))$ for the SDCM $\mathcal{R}$ in Example 3.3.8 need to satisfy $\overline{X}^{(1)}_{[0],0} = (0, 0)$ and $\overline{X}^{(1)}_{[0],d} = (L, 0)$ after the perfect intervention $\mathrm{do}(\{1, d\}, (0, L))$ on the model.

In summary, Examples 3.3.18, 3.3.20 and 3.3.21 show that the initial (random) value problems associated to dynamic SEs of an SDCM behave differently compared to those of RDEs, as not every initial condition is consistent, and the solutions may involve (higher-order) derivatives of the exogenous process $\boldsymbol{E}$.

3.3.6 *Existence and uniqueness of the solutions*

For RDEs, there exist sufficient conditions for the existence and uniqueness of the solutions with an initial condition, which are similar to the existence and uniqueness theorems for initial value problems for ODEs (Coddington and Levinson, 1955; Bunke, 1972; Kloeden and Platen, 1992). No similar theorem is known in such generality for dynamic SEs, although there are some weaker results of this type for differential-algebraic equations (Ascher and Petzold, 1998). In this subsection, we provide sufficient conditions for the existence and uniqueness of solutions with a specified initial condition, both locally (considering only a subset of the stochastic processes) and globally.

We start with an assumption on the form of the dynamic SEs for a subset of endogenous processses $\mathcal{O} \subseteq \mathcal{I}$. This assumption entails that for some subset $I \subseteq \mathcal{O}$, the dynamic SEs corresponding to I can be written as an RDE, while the remaining dynamic SEs for the complement $\mathcal{O} \setminus I$ can be solved uniquely for their corresponding endogenous processes in terms of the other processes appearing in these dynamic SEs. Additionally, smoothness conditions are imposed on exogenous processes and on dynamical causal mechanisms to ensure the required smoothness of the solution.[23]

Assumption 1-$(I \subseteq \mathcal{O})$**.** *For the SDCM $\mathcal{R}$ and subsets $I \subseteq \mathcal{O} \subseteq \mathcal{I}$, writing $J := \mathcal{O} \setminus I$ and $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$ with $\mathrm{col}(\mathcal{G}^a(\mathcal{R}))$ the "collapsed" graph,[24] the following both hold:*

1. *the order tuple $\boldsymbol{n}_I \geq 1$;*

23 The required smoothness of the solutions implies that we need to make assumptions about the smoothness of the exogenous processes and the dynamical causal mechanisms in the model. The assumption we made here is still rather crude in the sense that it suffices, but it is not at all necessary; if desired, one can arrive at weaker conditions by carefully tracing through the graph how the required smoothness of the solution can be guaranteed by demanding certain smoothness of each exogenous process and each dynamical causal mechanism individually.

24 We will abuse notation by using the notation $\mathrm{col}(\mathcal{G}^a(\mathcal{R}))$ for the graph that is isomorphic to the "collapsed" mixed graph of $\mathcal{G}^a(\mathcal{R})$ where the nodes are labeled by $\mathcal{I} \cup \mathcal{J}$ instead of $\{\tilde{i}^{(n_i)} : i \in \mathcal{I}\} \cup \{\{j\} : j \in \mathcal{J}\}$.

2. *there exist continuous functions* $\boldsymbol{g}_I : \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I} \times \boldsymbol{\mathcal{X}}_J \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \boldsymbol{\mathcal{X}}_I$ *and* $\boldsymbol{g}_J : \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I} \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \boldsymbol{\mathcal{X}}_J$ *such that*[25] *for all* $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ *and for all* $\bar{\boldsymbol{x}}^{(\boldsymbol{n})} \in \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1}$

$$\boldsymbol{x}_I^{(\boldsymbol{n}_I)} = \boldsymbol{g}_I(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{x}_J, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) \quad \iff \quad \boldsymbol{x}_I = \boldsymbol{f}_I(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e})$$

and

$$\boldsymbol{x}_J = \boldsymbol{g}_J(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) \quad \iff \quad \boldsymbol{x}_J = \boldsymbol{f}_J(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e})\,.$$

In particular, under Assumption 1-($\mathcal{I} \subseteq \mathcal{I}$) the dynamic structural equations of $\mathcal{R}$ are equivalent to an RDE. For an SDCM that satisfies Assumption 1-($I \subseteq \mathcal{I}$) with I a strict subset of $\mathcal{I}$, we can eliminate the processes $\boldsymbol{X}_{\mathcal{I}\setminus I}$ by substitution, giving an RDE for the endogenous processes I of the form

$$\boldsymbol{X}_I^{(\boldsymbol{n}_I)} = \boldsymbol{g}_I(\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_{\mathcal{I}\setminus I}(\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{E}), \boldsymbol{E})\,. \tag{3.3}$$

Every solution of the original SDCM satisfies this RDE, and every solution of this RDE induces a solution of the SDCM, if it is sufficiently smooth. For $\mathcal{O} \subsetneq \mathcal{I}$ we can think of Assumption 1-($I \subseteq \mathcal{O}$) as applying this assumption to the subsystem with endogenous processes $\mathcal{O}$, treating the remaining endogenous processes in $\mathcal{I} \setminus \mathcal{O}$ as external inputs of the subsystem. This will turn out to be useful in Section 3.3.7 for proving a Markov property.

Example 3.3.22. *Consider the price, supply and demand model of Example 3.3.15. This model satisfies Assumption 1-($I \subseteq \mathcal{O}$) for $I = \{P\}$ and $\mathcal{O} = \{S, P, D\}$. Substituting the zeroth-order dynamic structural equations into the first-order equation of X_P yields the RDE*

$$X_P' = \lambda(\beta_D - \beta_S)X_P + \lambda(E_D - E_S)\,. \tag{3.4}$$

If instead we take $\mathcal{O} = \{S, P\}$, then this yields the RDE

$$X_P' = \lambda(X_D - \beta_S X_P - E_S)\,,$$

where now X_D is treated as an external input of the subsystem $\mathcal{O}$.

We formalize the notions of the existence and uniqueness of solutions of a subsystem of the SDCM as follows.

Definition 3.3.23 (Unique solvability of an initial value problem)**.** *Consider an SDCM $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E}\rangle$, subsets $I \subseteq \mathcal{O} \subseteq \mathcal{I}$ such that $\boldsymbol{n}_I \geq 1$, $J := \mathcal{O} \setminus I$ and $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$. We call the* initial value problem $\langle \mathcal{R}, I, \mathcal{O}\rangle$ (uniquely) solvable *if for any partial initial condition $(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{n}_I-1)})$ and any $C^{\boldsymbol{n}_P}$-stochastic process $\boldsymbol{X}_P$, there exists an (a.s. unique) $C^{\boldsymbol{n}_\mathcal{O}}$-stochastic process $\boldsymbol{X}_\mathcal{O}$ that is a solution of the dynamic structural equations*[26]

$$\boldsymbol{X}_\mathcal{O} = \boldsymbol{f}_\mathcal{O}(\overline{\boldsymbol{X}}_\mathcal{O}^{(\boldsymbol{n}_\mathcal{O})}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P)\,,$$

25 For a subset $P \subseteq \mathcal{I} \cup \mathcal{J}$, we use the convention that we write $\overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}$ and $\boldsymbol{E}_P$ instead of $\overline{\boldsymbol{X}}_{P\cap\mathcal{I}}^{(\boldsymbol{n}_{P\cap\mathcal{I}})}$ and $\boldsymbol{E}_{P\cap\mathcal{J}}$ respectively, and adopt a similar notation for variables and their spaces.

26 These equations are equivalent to the dynamic structural equations (see Definition 3.3.12).

with partial initial condition $(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{n}_I-1)})$.

As a special case, we obtain the notion of unique solvability for SCMs (see also Section 2.3.2), where initial values play no role.

Definition 3.3.24 (Unique solvability of SCMs). *Let* $\mathcal{M} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \boldsymbol{E} \rangle$ *be an SCM,* $\mathcal{O} \subseteq \mathcal{I}$ *a subset and* $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$. *We say that* $\mathcal{M}$ *is uniquely solvable w.r.t.* $\mathcal{O}$ *if for any value* $\boldsymbol{x}_P \in \boldsymbol{\mathcal{X}}_P$ *and any value* $\boldsymbol{e}_P \in \boldsymbol{\mathcal{E}}_P$, *there exists an a.s. unique solution* $\boldsymbol{x}_{\mathcal{O}} \in \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ *of the structural equations*

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

Note that this corresponds to unique solvability of the initial value problem $\langle \mathcal{M}, \emptyset, \mathcal{O} \rangle$. Since SDCMs that satisfy Assumption 1-($I \subseteq \mathcal{O}$) have the property that they determine an RDE on the subset I, we can apply the existence and uniqueness results of RDEs on this subsystem, which leads to the following result.

Lemma 3.3.25. *Let* $\mathcal{R}$ *be an SDCM that satisfies Assumption 1-*($I \subseteq \mathcal{O}$) *for subsets* $I \subseteq \mathcal{O} \subseteq \mathcal{I}$. *Let* $J := \mathcal{O} \setminus I$ *and* $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$. *If the following three conditions hold:*

1. *the exogenous process* $\boldsymbol{E}_P$ *is continuous;*
2. *the composition of* $\boldsymbol{g}_I$ *with* $\boldsymbol{g}_J$ *is uniformly Lipschitz in its I-input, that is, there exists a constant*[27] $\kappa > 0$ *such that for all* $\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)} \in \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I}$, *for all* $\bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)} \in \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1}$ *and for all* $\boldsymbol{e}_P \in \boldsymbol{\mathcal{E}}_P$ *the condition*

$$\begin{aligned}\Big\| g_I\big(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P\big) \\ - g_I\big(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P\big) \Big\| \\ \leq \kappa \| x_I^{(0)} - y_I^{(0)} \|\,.\end{aligned}$$

   is satisfied, where $\|\cdot\|$ *denotes the Euclidean norm on* $\boldsymbol{\mathcal{X}}_I$;
3. *for each* $j \in J$, *either* $n_j = 0$, *or* g_j *only depends on* $\boldsymbol{e}_P$ *(that is,* $g_j(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) = \tilde{g}_j(\boldsymbol{e}_P)$ *for* $\tilde{g}_j : \boldsymbol{\mathcal{E}}_P \to \boldsymbol{\mathcal{X}}_j$*) and* $g_j(\boldsymbol{E}_P)$ *is a* C^{n_j}*-stochastic process;*

then $\langle \mathcal{R}, I, \mathcal{O} \rangle$ *is uniquely solvable.*

This lemma guarantees the existence and uniqueness of solutions for a large class of (subsystems of) SDCMs. Indeed, it states that for any partial initial condition $(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{n}_I-1)})$ and any $C^{\boldsymbol{n}_P}$-stochastic process $\boldsymbol{X}_P$ there exists an a.s. unique solution $\boldsymbol{X}_{\mathcal{O}}$ of the dynamic structural equations

$$\boldsymbol{X}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\overline{\boldsymbol{X}}_{\mathcal{O}}^{(\boldsymbol{n}_{\mathcal{O}})}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P)\,,$$

27 This result can be weakened slightly by making κ dependent on $t \in T$, $\omega \in \Omega$ and the parent processes $\overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P-1)}$ and $\boldsymbol{E}_P$ (see also Theorem 1.2 in Bunke (1972) or Theorem 3.2 in Neckel and Rupp (2013)).

with initial condition

$$\big(\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)}(t_0), \boldsymbol{X}_J(t_0)\big) = \Big(\overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J\big(\overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{n}_I-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}(t_0), \boldsymbol{E}_P(t_0)\big)\Big)$$

at t_0. In particular, this provides a sufficient condition for an initial condition to be consistent (see Definition 3.3.19).

In general, Assumption 1-($I \subseteq \mathcal{O}$) for an SDCM is not preserved under a stochastic perfect intervention. Consider for example the SDCM $\tilde{\mathcal{R}}$ in Example 3.3.10 which satisfies Assumption 1-($\mathcal{I} \subseteq \mathcal{I}$). Performing the intervention $\mathrm{do}(1, 0)$ on this model yields a model that does not satisfy Assumption 1-($I \subseteq \mathcal{I}$) for any $I \subseteq \mathcal{I}$. Under the following stronger assumption the SDCM will satisfy Assumption 1-($I \subseteq \mathcal{I}$) for some $I \subseteq \mathcal{I}$ after every stochastic perfect intervention.

Assumption 2-($I \subseteq \mathcal{O}$). *For the SDCM $\mathcal{R}$ and subsets $I \subseteq \mathcal{O} \subseteq \mathcal{I}$, writing $J := \mathcal{O} \setminus I$ and $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$, the following all hold:*

1. *the order tuple $\boldsymbol{n}_I \geq 1$;*
2. *there exist continuous functions $g_i : \mathcal{X}_i^{n_i} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}\setminus i} \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \mathcal{X}_i$ for all $i \in I$ and $g_j : \boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \mathcal{X}_j$ for all $j \in J$ such that for all $i \in I$, all $j \in J$, all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and all $\bar{\boldsymbol{x}}^{(\boldsymbol{n})} \in \boldsymbol{\mathcal{X}}^{\boldsymbol{n}+1}$,*

$$x_i^{(n_i)} = g_i(\bar{x}_i^{(n_i-1)}, \boldsymbol{x}_{\mathcal{O}\setminus i}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) \quad \iff \quad x_i = f_i(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e})$$

and

$$x_j = g_j(\boldsymbol{x}_I, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) \quad \iff \quad x_j = f_j(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e})\,.$$

In particular, Assumption 2-($I \subseteq \mathcal{O}$) implies Assumption 1-($I \subseteq \mathcal{O}$).

Proposition 3.3.26 (Assumption 2-($I \subseteq \mathcal{O}$) under stochastic perfect intervention)**.** *Let $\mathcal{R}$ be an SDCM that satisfies Assumption 2-($I \subseteq \mathcal{O}$) for subsets $I \subseteq \mathcal{O} \subseteq \mathcal{I}$. Then, for a stochastic perfect intervention $\mathrm{do}(L, \boldsymbol{K}_L)$ for $L \subseteq \mathcal{O}$, the intervened SDCM $\mathcal{R}_{\mathrm{do}(L,\boldsymbol{K}_L)}$ satisfies Assumption 2-($I \setminus L \subseteq \mathcal{O}$).*

This proposition shows the usefulness of Assumption 2-($I \subseteq \mathcal{O}$), in that it gives a guarantee that after any stochastic perfect intervention on a subset of $\mathcal{O}$, Assumption 1-($\tilde{I} \subseteq \mathcal{O}$) is satisfied for some $\tilde{I} \subseteq \mathcal{O}$, and hence Lemma 3.3.25 can be applied.

LINEAR SDCMS Observe that a linear SDCM that satisfies Assumption 1-($I \subseteq \mathcal{O}$) is of the following form.

Proposition 3.3.27. *Let $\mathcal{R}$ be a linear SDCM, $I \subseteq \mathcal{O} \subseteq \mathcal{I}$ be subsets, and let $J := \mathcal{O} \setminus I$ and $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$. Then $\mathcal{R}$ satisfies Assumption 1-($I \subseteq \mathcal{O}$) iff the dynamic causal mechanism $\boldsymbol{f}_{\mathcal{O}}$ of $\mathcal{R}$ restricted to $\mathcal{O}$ is of the form*

$$\begin{cases} \boldsymbol{f}_I(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := B_{II^{(\boldsymbol{n}_I)}} \boldsymbol{x}_I^{(\boldsymbol{n}_I)} + B_{I\bar{I}^{(\boldsymbol{n}_I-1)}} \bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)} + B_{IJ} \boldsymbol{x}_J + B_{I\bar{P}^{(\boldsymbol{n}_P)}} \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)} + \Gamma_{IP} \boldsymbol{e}_P \\ \boldsymbol{f}_J(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := B_{J\bar{I}^{(\boldsymbol{n}_I-1)}} \bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)} + B_{JJ} \boldsymbol{x}_J + \boldsymbol{x}_J + B_{J\bar{P}^{(\boldsymbol{n}_P)}} \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)} + \Gamma_{JP} \boldsymbol{e}_P\,, \end{cases}$$

where $B_{II^{(n_I)}}$ and B_{JJ} are invertible matrices.

In particular, for linear SDCMs, Lemma 3.3.25 gives the following useful corollary.

Corollary 3.3.28. *Let $\mathcal{R}$ be a linear SDCM, $I \subseteq \mathcal{O} \subseteq \mathcal{I}$ be subsets, and let $J := \mathcal{O} \setminus I$ and $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$. If*

1. *$\mathcal{R}$ satisfies Assumption 1-($I \subseteq \mathcal{O}$);*
2. *$\boldsymbol{E}_P$ is continuous;*
3. *for each $j \in J$, either $n_j = 0$, or $(B_{JJ}^{-1})_{jJ} B_{J\bar{I}^{(n_I - 1)}} = \mathbf{0}$, $(B_{JJ}^{-1})_{jJ} B_{J\bar{P}^{(n_P)}} = \mathbf{0}$ and $(B_{JJ}^{-1})_{jJ} \Gamma_{JP} \boldsymbol{E}_P$ is a C^{n_j}-stochastic process;*

then $\langle \mathcal{R}, I, \mathcal{O} \rangle$ is uniquely solvable.

Examples of linear SDCMs that satisfy Assumption 1-($I \subseteq \mathcal{I}$) for some subset I are the SDCMs $\mathcal{R}$ of Example 3.3.5 and $\mathcal{R}_{\mathrm{do}(\{1,d\},(0,L))}$ of Example 3.3.8, which satisfy Assumption 1-($\mathcal{I} \subseteq \mathcal{I}$) and 1-($\mathcal{I} \setminus \{1, d\} \subseteq \mathcal{I}$), respectively. As the other conditions in Corollary 3.3.28 are fulfilled, they both have an a.s. unique solution for each respective partial initial condition.

In particular, for linear SDCMs that satisfy Assumption 2-($I \subseteq \mathcal{O}$) we have the following corollary.

Corollary 3.3.29. *Let $\mathcal{R}$ be a linear SDCM, $I \subseteq \mathcal{O} \subseteq \mathcal{I}$ be subsets, and let $J := \mathcal{O} \setminus I$ and $P := \mathrm{pa}_{\mathrm{col}(\mathcal{G}^a(\mathcal{R}))}(\mathcal{O}) \setminus \mathcal{O}$. If*

1. *$\mathcal{R}$ satisfies Assumption 2-($I \subseteq \mathcal{O}$),*
2. *$\boldsymbol{E}_P$ is continuous;*
3. *for each $j \in J$, either $n_j = 0$, or $(B_{JJ}^{-1})_{jJ} B_{J\bar{I}^{(n_I - 1)}} = \mathbf{0}$, $(B_{JJ}^{-1})_{jJ} B_{J\bar{P}^{(n_P)}} = \mathbf{0}$ and $(B_{JJ}^{-1})_{jJ} \Gamma_{JP} \boldsymbol{E}_P$ is a C^{n_j}-stochastic process;*

then $\langle \mathcal{R}_{\mathrm{do}(L, \boldsymbol{K}_L)}, I \setminus L, \mathcal{O} \rangle$ is uniquely solvable for any stochastic perfect intervention $\mathrm{do}(L, \boldsymbol{K}_L)$ with $L \subseteq \mathcal{O}$ and $\boldsymbol{K}_L$ a $C^{\boldsymbol{n}_L}$-stochastic process.

Examples of linear SDCMs that satisfy Assumption 2-($I \subseteq \mathcal{I}$) for some subset I are the damped coupled harmonic oscillator of Example 3.3.5 and the price, supply and demand model of Example 3.3.15. Hence, the existence of solutions is guaranteed for both models after any (sufficiently smooth) stochastic perfect intervention, and the solutions are a.s. uniquely determined by the respective partial initial conditions.

NONLINEAR SDCMS An example of an SDCM that is not linear but satisfies Assumption 2-($I \subseteq \mathcal{O}$) is the bathtub model discussed in (Iwasaki and Simon, 1994). The existence and uniqueness conditions apply to this particular model.

Example 3.3.30 (Bathtub model). *Water enters a bathtub from the faucet at a certain rate X_{Q_i} and exits the bathtub via the drain at a rate X_{Q_o}. The drain has a diameter of X_K, the depth of the water is X_D and the pressure at the base of the drain is X_P. Iwasaki and Simon (1994) propose to model this as a dynamical system with (random) differential equations given by*

$$\begin{cases} X_K = k_0 \\ X_{Q_i} = q_0 \\ X'_P = \alpha_2(\alpha_4 X_D - X_P) \\ X'_{Q_o} = \alpha_3(\alpha_1 X_K X_P - X_{Q_o}) \\ X'_D = \alpha_0(X_{Q_i} - X_{Q_o})\,, \end{cases} \tag{3.5}$$

where $k_0, q_0 \in \mathbb{R}_{>0}$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_4) \in \mathbb{R}^5_{>0}$ are some constants. We consider the dynamic causal mechanism

$$\begin{cases} f_K(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := e_K \\ f_{Q_i}(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := e_{Q_i} \\ f_P(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := \alpha_4 x_D - \alpha_2^{-1} x_{P'} \\ f_{Q_o}(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := \alpha_1 x_K x_P - \alpha_3^{-1} x_{Q'_o} \\ f_D(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := x_D + \alpha_0(x_{Q_i} - x_{Q_o}) - x_{D'}\,. \end{cases}$$

with order tuple $\boldsymbol{n} := (n_K, n_{Q_i}, n_P, n_{Q_o}, n_D) = (0, 0, 1, 1, 1)$ and the exogenous processes are given by $E_K(t, \omega) := k_0$, $E_{Q_i}(t, \omega) := q_0$. The dynamic structural equations of this SDCM, denoted by $\mathcal{R}$, read

$$\begin{cases} X_K = E_K \\ X_{Q_i} = E_{Q_i} \\ X_P = \alpha_4 X_D - \alpha_2^{-1} X'_P \\ X_{Q_o} = \alpha_1 X_K X_P - \alpha_3^{-1} X'_{Q_o} \\ X_D = X_D + \alpha_0(X_{Q_i} - X_{Q_o}) - X'_D\,, \end{cases}$$

and have the same solutions as the system of equations (3.5) *(see also Footnote 10). The corresponding SDCM graph is depicted in Figure 3.7 (top left). This SDCM of the bathtub model satisfies Assumption 2-($\{P, Q_o, D\} \subseteq \mathcal{I}$) with $\mathcal{I} = \{K, Q_i, P, Q_o, D\}$, and hence, after any sufficiently smooth stochastic perfect intervention $\mathrm{do}(L, \boldsymbol{K}_L)$ with $L \subseteq \mathcal{I}$, the intervened bathtub model $\mathcal{R}_{\mathrm{do}(L,\boldsymbol{K}_L)}$ satisfies Assumption 2-($\{P, Q_o, D\} \setminus L \subseteq \mathcal{I}$). Since the induced RDE of the intervened model $\mathcal{R}_{\mathrm{do}(L,\boldsymbol{K}_L)}$ on the endogenous processes $\{P, Q_o, D\} \setminus L$ is linear in these endogenous processes, it follows from Lemma 3.3.25 that (for sufficiently smooth exogenous process $\boldsymbol{K}_L$) $\mathcal{R}_{\mathrm{do}(L,\boldsymbol{K}_L)}$ has an a.s. unique solution for any partial initial condition $(t_0, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{\{P,Q_o,D\}\setminus L})}_{\{P,Q_o,D\}\setminus L,[0]})$.*

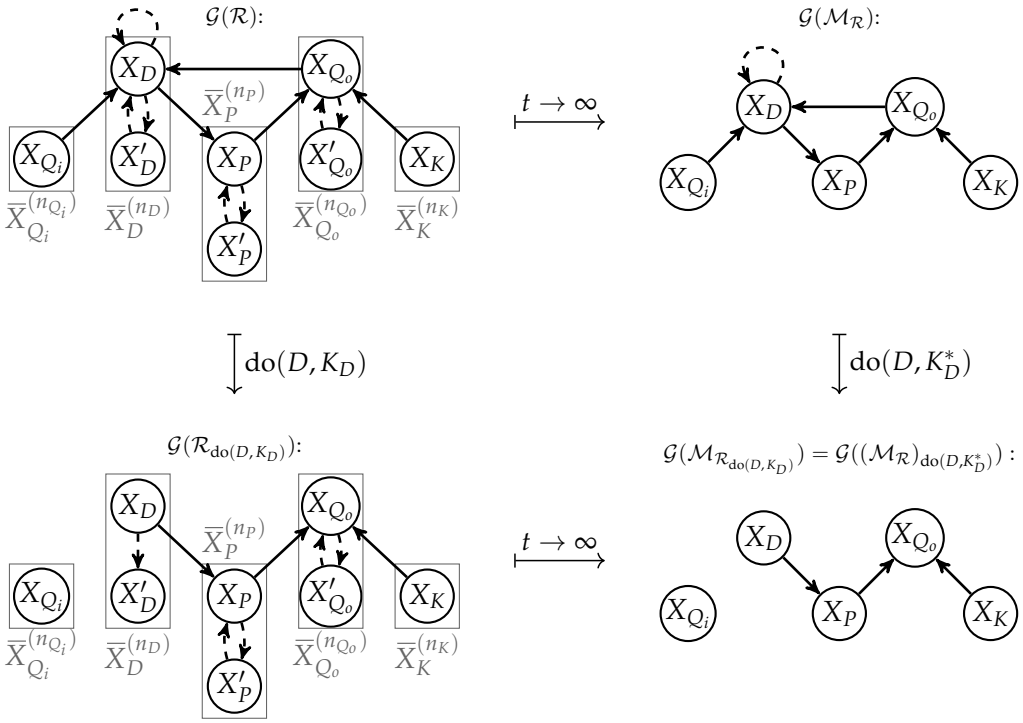


Figure 3.7: Graphs of the bathtub model: original model $\mathcal{R}$ of Example 3.3.30 (top left), the equilibrated model $\mathcal{M}_{\mathcal{R}}$ (top right), the intervened model $\mathcal{R}_{\mathrm{do}(D,K_D)}$ (bottom left), and the intervened and equilibrated model $\mathcal{M}_{\mathcal{R}_{\mathrm{do}(D,K_D)}}$ (bottom right) of Example 3.4.19.

3.3.7 *Markov property for SDCMs with initial conditions*

Theoretical results of key importance concerning SCMs are their so-called Markov properties, which allow to read off conditional independencies in the solutions of an SCM from the graph of the SCM (see Section 2.6 and Appendix 2.A.2). The two most well-known Markov properties for SCMs are the d-separation criterion (which applies to acyclic SCMs, amongst others), and the σ-separation criterion (which applies for example to the more general class of simple SCMs that can contain causal cycles). Here we derive a Markov property for SDCMs with initial conditions that is analogous to the σ-separation criterion for SCMs.

Key to proving Markov properties is the existence and uniqueness of solutions for each subsystem consisting of one strongly connected component of the collapsed graph of the SDCM, augmented with initial conditions. By reinterpreting continuous stochastic processes as random variables taking values in a space of continuous functions, we can make use of the existing σ-separation Markov property for SCMs to derive Markov properties for SDCMs.

To avoid complicating matters further with smoothness assumptions, we will assume that the order tuple is as small as possible.

Definition 3.3.31 (Tight order tuple)**.** *Let $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ be an SDCM. Its order tuple $\boldsymbol{n}$ is called* tight *if for each $i \in \mathcal{I}$, either $n_i = 0$, or $n_i > 0$ and the edge $i^{(n_i)} \to i^{(0)}$ appears in $\mathcal{G}^a(\mathcal{R})$.*

Note that the order tuple is tight if and only if each cluster $\bar{i}^{(n_i)}$ in the augmented graph $\mathcal{G}^a(\mathcal{R})$ forms a cycle in the cluster, that is, if there is a directed path in the cluster from each node in the cluster to any other node in the cluster.

Definition 3.3.32 (Augmented collapsed graph for SDCMs)**.** *Consider an SDCM $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ with tight order tuple. We define the* augmented collapsed graph $\mathcal{G}^+_{[0]}(\mathcal{R})$ *of $\mathcal{R}$ as the directed graph with nodes $\mathcal{I} \cup \mathcal{J} \cup \mathcal{I}_{[0]}$, where $\mathcal{I}_{[0]} := \{i_{[0]} : i \in \mathcal{I} : n_i \geq 1\}$, directed edges $k \to i$ (but dashed $i \dashrightarrow i$ if $k = i$) if either $k^{(m_k)} \in \overline{\mathcal{I}}^{(\boldsymbol{n})}$ is functional parent of $i \in \mathcal{I}$ for some m_k or $k \in \mathcal{J}$ is functional parent of $i \in \mathcal{I}$, and additional directed edges $i_{[0]} \to i$ for those $i \in \mathcal{I}$ with $i_{[0]} \in \mathcal{I}_{[0]}$.*

The nodes $i_{[0]}$ represent partial initial conditions $(t_0, \overline{X}^{(n_i-1)}_{[0],i})$, while the nodes in $\mathcal{I} \cup \mathcal{J}$ represent endogenous stochastic processes X_i for $i \in \mathcal{I}$, and exogenous stochastic processes E_j for $j \in \mathcal{J}$. The augmented collapsed graph of an SDCM (with tight order tuple) is similar to its augmented graph, except that clusters are collapsed and nodes representing initial conditions have been added. Figure 3.8 (top right) shows the augmented collapsed graph for the bathtub model of Example 3.3.30, and for comparison, the augmented graph is also shown (top left).

We can now prove that under conditions that guarantee the existence and uniqueness of a solution locally for each strongly connected component of the augmented collapsed graph, there exists a global solution that is unique and satisfies the σ-separation criterion with respect to the augmented collapsed graph.

Theorem 3.3.33 (Markov property for SDCMs with initial conditions)**.** *Consider an SDCM $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ with tight order tuple. Suppose that for each strongly connected component $S \subseteq \mathcal{I}$ of $\mathcal{G}^+_{[0]}(\mathcal{R})$, $\mathcal{R}$ satisfies Assumption 1-$(I_S \subseteq S)$ for some subset $I_S \subseteq S$ and $\langle \mathcal{R}, I_S, S \rangle$ is uniquely solvable. Then for any partial initial condition $\big(t_0, (\overline{X}^{(n_i-1)}_{[0],i})_{i \in \mathcal{I}_{[0]}}\big)$, the SDCM $\mathcal{R}$ has an a.s. unique solution with that partial initial condition. If $(\overline{X}^{(n_i-1)}_{[0],i})_{i \in \mathcal{I}_{[0]}}$ is independent, and independent of $\boldsymbol{E}$, the solution $\boldsymbol{X}$ satisfies the following Markov property:*

$$A \overset{\sigma}{\underset{\mathcal{G}^+_{[0]}(\mathcal{R})}{\perp}} B \,|\, C \implies \boldsymbol{Z}_A \perp\!\!\!\perp \boldsymbol{Z}_B \,|\, \boldsymbol{Z}_C$$

for all subsets of nodes A, B, C of $\mathcal{G}^+_{[0]}(\mathcal{R})$, where $\boldsymbol{Z}_A := (\boldsymbol{X}_{A \cap \mathcal{I}}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{A \cap \mathcal{I}_{[0]}})}_{[0],A \cap \mathcal{I}_{[0]}}, \boldsymbol{E}_{A \cap \mathcal{J}})$ for $A \subseteq \mathcal{I} \cup \mathcal{I}_{[0]} \cup \mathcal{J}$.

The conditional independence in this Markov property requires to interpret the endogenous process $\boldsymbol{X}$ as a random element of $\mathcal{C}^{\boldsymbol{n}}(T, \boldsymbol{\mathcal{X}})$ and the exogenous process $\boldsymbol{E}$ as a random element of $\mathcal{C}^0(T, \boldsymbol{\mathcal{E}})$. In other words, we may conclude the independence of entire processes (and initial conditions), conditional on entire processes (and initial conditions).

We can extend this result to obtain a Markov property for the solutions evaluated at times t_0 and t_1. For this, we extend the graph with nodes that correspond to evaluating the endogenous processes at time t_1.

Definition 3.3.34 (Evaluated augmented collapsed graph for SDCMs)**.** *Consider an SDCM $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ with tight order tuple. We define the* evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$ *of $\mathcal{R}$ as the augmented collapsed graph $\mathcal{G}^+_{[0]}$, extended with additional nodes $\mathcal{I}_{[1]} := \{i_{[1]} : i \in \mathcal{I}\}$ and directed edges $i \to i_{[1]}$ for $i \in \mathcal{I}$.*

The additional nodes in the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$ correspond with the evaluation of a process at time t_1, that is, $i_{[1]}$ corresponds with $\overline{X}_i^{(n_i)}(t_1)$. Figure 3.8 (bottom left) shows the evaluated augmented collapsed graph for the bathtub model of Example 3.3.30. We get the following corollary almost for free.

Corollary 3.3.35. *Under the assumptions of Theorem 3.3.33, the following Markov property also holds:*

$$A \overset{\sigma}{\underset{\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})}{\perp}} B \,|\, C \implies \boldsymbol{Z}_A \perp\!\!\!\perp \boldsymbol{Z}_B \,|\, \boldsymbol{Z}_C$$

for any subsets of nodes A, B, C of the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$, where for $A \subseteq \mathcal{I} \cup \mathcal{I}_{[0]} \cup \mathcal{I}_{[1]} \cup \mathcal{J}$ we write $\boldsymbol{Z}_A := (\boldsymbol{X}_{A\cap\mathcal{I}}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{A\cap\mathcal{I}_{[0]}})}_{[0],A\cap\mathcal{I}_{[0]}}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{A\cap\mathcal{I}_{[1]}})}_{A\cap\mathcal{I}_{[1]}}(t_1), \boldsymbol{E}_{A\cap\mathcal{J}})$ with $\boldsymbol{X}$ being an a.s. unique solution of $\mathcal{R}$ with initial condition $\big(t_0, (\overline{X}^{(n_i-1)}_{[0],i})_{i\in\mathcal{I}_{[0]}}\big)$.

We can also marginalize out the "process nodes" and retain only the "random variable" nodes, in effect only considering observations of the processes at times t_0 and t_1.

Definition 3.3.36 (Transition graph for SDCMs)**.** *Let $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ be an SDCM with tight order tuple. We define the* transition graph $\mathcal{G}_{[0]\dots[1]}(\mathcal{R})$ *of $\mathcal{R}$ as the directed graph with nodes $\mathcal{I}_{[0]} \cup \mathcal{I}_{[1]} \cup \mathcal{J}$, where $\mathcal{I}_{[0]} := \{i_{[0]} : i \in \mathcal{I} : n_i \geq 1\}$ and $\mathcal{I}_{[1]} := \{i_{[1]} : i \in \mathcal{I}\}$, and directed edges $i \to j$ if there exists a directed path $i \to \dots \to j$ in the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$.*

The transition graph $\mathcal{G}_{[0]\dots[1]}(\mathcal{R})$ is obtained from the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$ by graphically marginalizing[28] out the nodes $\mathcal{I}$ representing the full endogenous processes, and keeping only the nodes $\mathcal{I}_{[0]} \cup \mathcal{I}_{[1]}$ corresponding with the evaluations of the processes at time t_0 and time t_1, in addition to the nodes $\mathcal{J}$ corresponding with the exogenous processes. Figure 3.8 (bottom right) shows the transition graph for the bathtub model of Example 3.3.30.

Corollary 3.3.37. *Under the assumptions of Theorem 3.3.33, the following Markov property also holds:*

$$A \overset{\sigma}{\underset{\mathcal{G}_{[0]\dots[1]}(\mathcal{R})}{\perp}} B \,|\, C \implies \boldsymbol{Z}_A \perp\!\!\!\perp \boldsymbol{Z}_B \,|\, \boldsymbol{Z}_C$$

28 The result of a graphical marginalization is also known as the "latent projection" (see Definition 2.5.7).

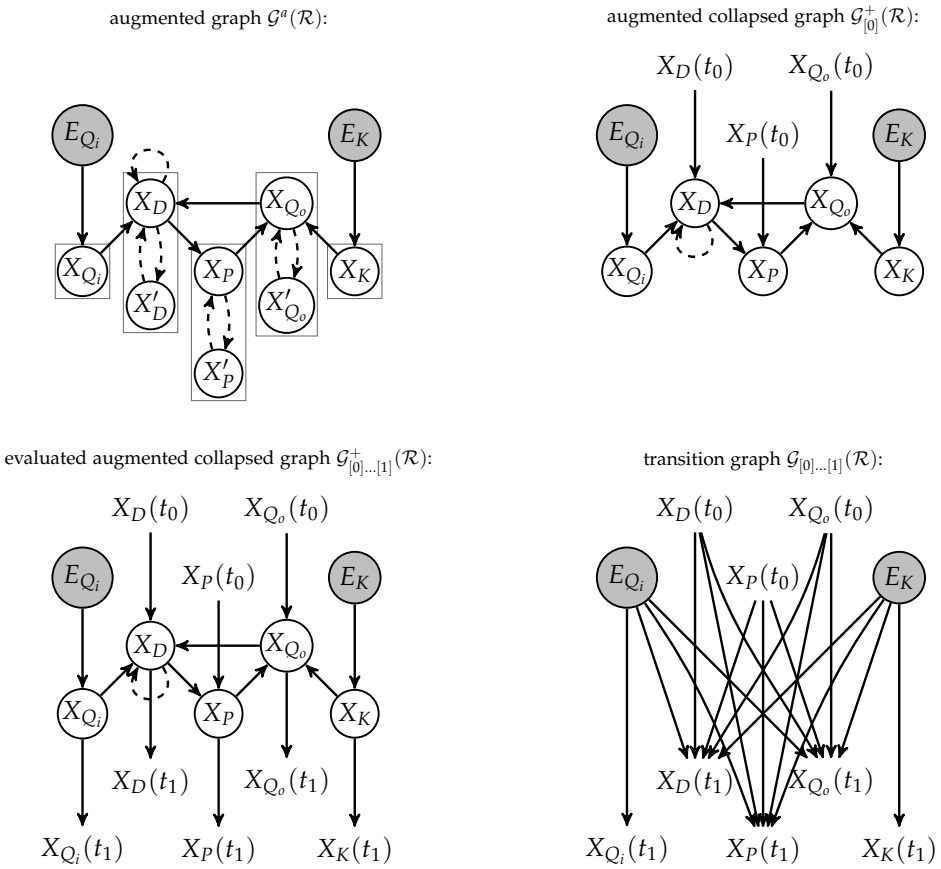


Figure 3.8: Different graphs of the bathtub model of Examples 3.3.30 and 3.3.38.

for any subsets of nodes A, B, C of the transition graph $\mathcal{G}_{[0]\dots[1]}(\mathcal{R})$, where for $A \subseteq \mathcal{I}_{[0]} \cup \mathcal{I}_{[1]} \cup \mathcal{J}$ we write $\boldsymbol{Z}_A := (\overline{\boldsymbol{X}}^{(\boldsymbol{n}_{A\cap\mathcal{I}_{[0]}})}_{[0],A\cap\mathcal{I}_{[0]}}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{A\cap\mathcal{I}_{[1]}})}_{A\cap\mathcal{I}_{[1]}}(t_1), \boldsymbol{E}_{A\cap\mathcal{J}})$ with $\boldsymbol{X}$ being an a.s. unique solution of $\mathcal{R}$ with initial condition $\big(t_0, (\overline{X}^{(n_i-1)}_{[0],i})_{i\in\mathcal{I}_{[0]}}\big)$.

Example 3.3.38 (Markov properties for the bathtub model)**.** *The bathtub model of Example 3.3.30 satisfies the assumptions of Theorem 3.3.33 and its Corollaries 3.3.35 and 3.3.37. The corresponding graphs are illustrated in Figure 3.8. We can, for example, read off from the augmented collapsed graph $\mathcal{G}^+_{[0]}(\mathcal{R})$ that $X_{Q_i} \perp\!\!\!\perp X_K$. From the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$ and the transition graph $\mathcal{G}_{[0]\dots[1]}(\mathcal{R})$ we can read off that $X_K(t_1) \perp\!\!\!\perp X_{Q_i}(t_1)$, that is, the inflow through the faucet is independent of the drain diameter at time t_1, provided they are at t_0. The latter is hardly surprising, but serves to illustrate how one can use the Markov properties to arrive at conditional independence statements about the solution without actually solving the SDCM, by carefully tracing the functional relations encoded in the dynamic structural equations of the model.*

3.4 EQUILIBRATION OF SDCMS

In this section, we will take $T = [t_0, \infty)$ and study the equilibrium states of SDCMs and, in particular, of steady SDCMs, which are SDCMs for which the dynamic

structural equations and exogenous processes become explicitly time-independent asymptotically as $t \to \infty$. We introduce an equilibration operation on a steady SDCM, which equilibrates the model to an SCM such that all the equilibrium states of the SDCM are described by the solutions of the SCM. Intuitively, this equilibration operation separately equilibrates each dynamic causal mechanism, which corresponds mathematically to transforming each dynamic structural equation into a structural equation of the SCM. We show that this equilibration operation commutes with perfect stochastic interventions, without requiring the strong global stability assumption of Mooij, Janzing, and Schölkopf (2013), which assumes that all the solutions equilibrate to the same static equilibrium state. This allows to study the causal semantics of the equilibrium states of steady SDCMs within the framework of SCMs.

We start in Section 3.4.1 with the definition of equilibrating solutions and their corresponding equilibrium states. In Section 3.4.2, we define the class of steady SDCMs which have several convenient convergence properties. In Section 3.4.3, we show how one can equilibrate a steady SDCM to an SCM. In Section 3.4.4, we show how the equilibration acts on the graph of an SDCM. In Section 3.4.5, we show that the equilibration operation commutes with intervention. We discuss in Section 3.4.6 the inverse problem of finding steady SDCMs for which all the solutions equilibrate to solutions of the SCM independently of the initial condition. We provide sufficient conditions under which one can construct a first-order steady SDCM such that its equilibration coincides with a given linear SCM. This establishes a class of linear SCMs that model the causal equilibrium semantics of certain linear dynamical systems. In Section 3.4.7, we discuss some subtleties in the causal interpretation of the graph of the equilibrated SDCM.

3.4.1 *Equilibrating solutions and equilibrium states*

In this subsection, we define the equilibrating solutions of an SDCM as those solutions for which all the higher-order derivatives that are considered in the model converge to zero a.s.. For a stochastic process $\boldsymbol{X}$ we say that it *converges almost surely* to a random variable $\boldsymbol{X}^*$, if the limit $\lim_{t\to\infty} \boldsymbol{X}_t$ exists almost surely[29] and is a.s. equal to $\boldsymbol{X}^*$. In this case, we call $\boldsymbol{X}$ *almost surely convergent*.

Definition 3.4.1 (Equilibrating solution, equilibrium state)**.** *Let $\boldsymbol{X}$ be a solution of an SDCM $\mathcal{R}$. We call $\boldsymbol{X}$ an* equilibrating solution*, if $\overline{\boldsymbol{X}}^{(n)}$ is a.s. convergent. In particular, an equilibrating solution $\boldsymbol{X}$ converges almost surely to a random variable $\boldsymbol{X}^*$, and we say that $\boldsymbol{X}$* equilibrates *to $\boldsymbol{X}^*$ and call $\boldsymbol{X}^*$ an* equilibrium state *of $\mathcal{R}$.*

An example of an SDCM with equilibrium states is the price, supply and demand model of Example 3.3.15, where the equilibrium states correspond to "market equilibrium", as illustrated in the following example.

29 In that case, it defines a random variable, because $\lim_{t\to\infty} \boldsymbol{X}_t = \lim_{\substack{t\to\infty \\ t\in\mathbb{N}}} \boldsymbol{X}_t$ a.s., and the latter is a random variable.

Example 3.4.2 (Market equilibrium)**.** *Consider the price, supply and demand model of Example 3.3.15 with E_S and E_D constant exogenous processes. Market equilibrium for this model is reached if*

$$X_D^* - X_S^* = 0\,,$$

that is, if the demanded and supplied quantities become equal asymptotically. The solutions that satisfy this condition are equilibrating solutions for which

$$X_P'^* = 0\,, \qquad X_P^* = \frac{E_D - E_S}{\beta_S - \beta_D}\,, \quad X_S^* = X_D^* = \frac{\beta_S E_D - \beta_D E_S}{\beta_S - \beta_D}\,.$$

In fact, for every solution $\boldsymbol{X}$ that equilibrates, the higher-order derivatives of $\boldsymbol{X}$ must converge to zero almost surely.

Proposition 3.4.3. *Let $\boldsymbol{X}$ be a solution of an SDCM $\mathcal{R}$. If $\boldsymbol{X}$ equilibrates, then $\lim_{t\to\infty} \overline{X}_i^{(n_i)} = (X_i^*, 0, \ldots, 0)$ a.s. for all $i \in \mathcal{I}$, where X_i^* is the i^{th} component of the corresponding equilibrium state $\boldsymbol{X}^*$.*

In particular, for linear SDCMs we can show that all the solutions of the SDCM equilibrate under certain conditions.

Proposition 3.4.4. *Let $\mathcal{R}$ be a linear SDCM that satisfies Assumption 1-($I \subseteq \mathcal{I}$) for a subset $I \subseteq \mathcal{I}$ with an order tuple $\boldsymbol{n}_I = 1$ and an exogenous process $\boldsymbol{E}$ that is constant in time.[30] By Proposition 3.3.27, the dynamical causal mechanism $\boldsymbol{f}$ is of the form*

$$\begin{cases} \boldsymbol{f}_I(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := B_{II'}\boldsymbol{x}_I' + B_{II}\boldsymbol{x}_I + B_{IJ}\boldsymbol{x}_J + \Gamma_{I\mathcal{J}}\boldsymbol{e} \\ \boldsymbol{f}_J(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, \boldsymbol{e}) := B_{JI}\boldsymbol{x}_I + B_{JJ}\boldsymbol{x}_J + \boldsymbol{x}_J + \Gamma_{J\mathcal{J}}\boldsymbol{e}\,, \end{cases}$$

where $J := \mathcal{I} \setminus I$ and $B_{II'}$ and B_{JJ} are invertible matrices. If the matrix $B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}B_{JI} - B_{II} + \mathbb{I}_I)$, where $\mathbb{I}_I$ denotes the identity matrix, is Hurwitz (that is, every eigenvalue has a strictly negative real part), then every solution $\boldsymbol{X}$ of $\mathcal{R}$ equilibrates to the same equilibrium state, irrespective of the initial condition.

This proposition allows us to derive a condition for which the price, supply and demand model always reaches market equilibrium.

Example 3.4.5 (Market equilibrium, continued)**.** *Applying Proposition 3.4.4 to the price, supply and demand model of Example 3.3.15 shows that $B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}B_{JI} - B_{II} + \mathbb{I}_I) = \lambda(\beta_D - \beta_S)$. This matrix is Hurwitz if and only if $\lambda(\beta_D - \beta_S) < 0$. Thus, since $\lambda > 0$, the price X_P, supply X_S and demand X_D equilibrate for constant exogenous processes E_D and E_S if $\beta_S > \beta_D$.*

30 In general, we can let $\boldsymbol{E}$ be a continuous exogenous process that depends on time as long as both $\boldsymbol{E}_t$ and $\exp(At)\int_{t_0}^t \exp(-As)C\boldsymbol{E}(s)ds$ converge almost surely for $t \to \infty$, where $A := B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}B_{JI} - B_{II} + \mathbb{I}_I)$ and $C := B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}\Gamma_{J\mathcal{J}} - \Gamma_{I\mathcal{J}})$. In that case, the order tuple may matter, and it must be checked whether the solutions are sufficiently smooth.

3.4.2 *Steady SDCMs*

In this subsection, we define the class of steady SDCMs which have the convenient property that their dynamics become explicitly time-independent asymptotically for $t \to \infty$.

Definition 3.4.6 (Steady SDCM)**.** *We call an SDCM $\mathcal{R}$* steady*, if it has a dynamic causal mechanism $\boldsymbol{f}$ that is continuous and an exogenous process $\boldsymbol{E}$ that is a.s. convergent.*

The continuity of the dynamic causal mechanism and the convergence assumption on the exogenous process assure us that the equilibrium states satisfy asymptotic dynamic structural equations.

Lemma 3.4.7. *Let $\mathcal{R}$ be a steady SDCM and let $\boldsymbol{E}^*$ be the random variable to which the exogenous process $\boldsymbol{E}$ converges a.s.. If $\boldsymbol{X}$ is an equilibrating solution of a steady SDCM $\mathcal{R}$, then the random variable $\overline{\boldsymbol{X}}^{(\boldsymbol{n})*}$ to which the complete $\boldsymbol{n}^{th}$-order derivative $\overline{\boldsymbol{X}}^{(\boldsymbol{n})}$ converges satisfies*

$$\boldsymbol{X}^* = \boldsymbol{f}(\overline{\boldsymbol{X}}^{(\boldsymbol{n})*}, \boldsymbol{E}^*) \quad a.s..$$

In general, not all solutions of a steady SDCM have to be equilibrating solutions, as one sees for example in Example 3.3.18.

The class of steady SDCMs is not closed under stochastic perfect interventions, since performing a stochastic perfect intervention that is not a.s. convergent yields an SDCM that is not steady. However, the class of steady SDCMs is closed under the following class of interventions.

Definition 3.4.8 (Steady stochastic perfect intervention)**.** *We call a stochastic perfect intervention* $\mathrm{do}(I, \boldsymbol{K}_I)$ *a* steady stochastic perfect intervention *if the process $\boldsymbol{K}_I$ converges a.s. to a random variable $\boldsymbol{K}_I^*$. We call it a* steady perfect intervention *if in addition $\boldsymbol{K}_I^* \in \boldsymbol{\mathcal{X}}_I$ (that is, it does not depend on ω).*

3.4.3 *Equilibration of a steady SDCM*

In this subsection, we show how we can equilibrate a steady SDCM to an SCM, such that the equilibrium states of the SDCM are described by the SCM. In the previous subsections, we saw that for an equilibrating solution of a steady SDCM, all the higher-order derivatives converge to zero, and the corresponding equilibrium state satisfies the asymptotic dynamic structural equations. Hence, we can construct an SCM from a steady SDCM such that every equilibrium state of the steady SDCM is a solution of this SCM.

Definition 3.4.9 (Equilibration of an SDCM)**.** *Let $\mathcal{R} = \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{n}, \boldsymbol{f}, \boldsymbol{E} \rangle$ be a steady SDCM and let $\boldsymbol{E}^*$ be a random variable such that $\boldsymbol{E}$ converges a.s. to it. We call the SCM $\mathcal{M}_{\mathcal{R}} := \langle \mathcal{I}, \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}^*, \boldsymbol{E}^* \rangle$ with the* equilibrated dynamic causal mechanism *$\boldsymbol{f}^* : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ given by*

$$\boldsymbol{f}^*(\boldsymbol{x}, \boldsymbol{e}) := \boldsymbol{f}(\bar{\boldsymbol{\iota}}(\boldsymbol{x}), \boldsymbol{e}),$$

an equilibration *of $\mathcal{R}$, where the mapping $\bar{\boldsymbol{\iota}} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}^{n+1}$ defined by*

$$\bar{\iota}_i^{(k_i)}(\boldsymbol{x}) = \begin{cases} x_i & \text{if } k_i = 0 \\ 0 & \text{otherwise,} \end{cases}$$

is the embedding that sets all the higher-order derivatives of the endogenous processes to 0.

In other words, the equilibration of an SDCM sets all the higher-order derivative entries in its dynamic causal mechanism to zero and replaces its exogenous process by its limiting random variable. In particular, linearity is preserved under equilibration, that is, a steady linear SDCM equilibrates to a linear SCM.

The equilibration of an SDCM is well defined due to the following result, which shows that the independence property for the family of exogenous processes $(\boldsymbol{E}_j)_{j\in\mathcal{J}}$ is preserved in the limit when time tends to infinity.

Proposition 3.4.10. *Let $(\boldsymbol{E}_j)_{j\in\mathcal{J}}$ be a family of stochastic processes, where $\mathcal{J}$ is some finite index set, such that $\boldsymbol{E}_j$ converges almost surely to the random variable $\boldsymbol{E}_j^*$, for every $j \in \mathcal{J}$. Then, if $(\boldsymbol{E}_j)_{j\in\mathcal{J}}$ is independent, so is the family of random variables $(\boldsymbol{E}_j^*)_{j\in\mathcal{J}}$.*

This equilibration of an SDCM to an SCM leads to the main insight that SCMs are capable of modeling all the equilibrium states of steady SDCMs.

Theorem 3.4.11. *If $\boldsymbol{X}$ is an equilibrating solution of a steady SDCM $\mathcal{R}$, then its limit $\boldsymbol{X}^*$ is a solution of the corresponding equilibration $\mathcal{M}_{\mathcal{R}}$.*

Intuitively, the equilibration of a steady SDCM to an SCM can be seen as the approximation of the dynamic structural equations by the structural equations of the SCM, which becomes exact at equilibrium. This is illustrated in the following example.

Example 3.4.12 (Equilibrated damped coupled harmonic oscillator)**.** *Consider the intervened damped coupled harmonic oscillator of Example 3.3.8 for which the dynamic structural equations are specified by*

$$\begin{cases} X_1 = 0 \\ X_i = \dfrac{\kappa_i}{\kappa_i + \kappa_{i-1}}(X_{i+1} - L_i) + \dfrac{\kappa_{i-1}}{\kappa_i + \kappa_{i-1}}(X_{i-1} + L_{i-1}) \\ \qquad\quad - \dfrac{b_i}{\kappa_i + \kappa_{i-1}} X_i' - \dfrac{m_i}{\kappa_i + \kappa_{i-1}} X_i'' \qquad (i = 2, \ldots, d-1) \\ X_d = L\,, \end{cases}$$

and where the exogenous processes $\boldsymbol{L}$ are random variables. In the limit, as time tends to infinity, the equilibrating solutions of the SDCM converge to the equilibrium states of the

equilibrated SDCM, which can be obtained by setting the higher-order derivatives to zero. This yields the equations

$$\begin{cases} X_1^* = 0 \\ X_i^* = \dfrac{\kappa_i(X_{i+1}^* - L_i) + \kappa_{i-1}(X_{i-1}^* + L_{i-1})}{\kappa_i + \kappa_{i-1}} & (i = 2, \dots, d-1) \\ X_d^* = L\,, \end{cases}$$

which describe the equilibrium states for the positions of the masses. Not all solutions necessarily equilibrate to an equilibrium, which happens for example in the case when there is no friction, that is, $b_i = 0$ for all $i \in \{2, \dots, d-1\}$. In this case, if any mass m_i starts at an off-equilibrium position (that is, if $X_i'(t_0) \neq 0$ or $X_i(t_0) \neq X_i^*$ for some $i \in \{2, \dots, d-1\}$), the solution will not equilibrate, but will keep on oscillating forever.

In case there is friction and the exogenous processes $\boldsymbol{L}$ are fixed to constant values, the equilibrated damped coupled harmonic oscillator exactly coincides with the deterministic SCM derived in (Mooij, Janzing, and Schölkopf, 2013). In Section 3.4.5 we will show that the equilibration operation, as defined in Definition 3.4.9, also preserves the causal semantics. The next example illustrates that our equilibration operation can also be applied to models that cannot be treated with the theory of (Mooij, Janzing, and Schölkopf, 2013).

Example 3.4.13 (Equilibrated price, supply and demand model)**.** *Setting the higher-order derivatives of the price, supply and demand model $\mathcal{R}$ of Example 3.3.15 to zero yields the structural equations:*

$$\begin{cases} X_P^* = X_P^* + \lambda(X_D^* - X_S^*) \\ X_S^* = \beta_S X_P^* + E_S^* \\ X_D^* = \beta_D X_P^* + E_D^*\,. \end{cases}$$

The equations describe the market equilibrium states. In Figure 3.9, we simulate the solutions of the SDCM $\mathcal{R}$ for random constant exogenous influences E_S and E_D and random consistent initial conditions. The dispersion of X_P, X_S and X_D at large t illustrates that the equilibrium state is not unique and depends on the initial condition. Hence, this example cannot be treated with the theory of (Mooij, Janzing, and Schölkopf, 2013).

Richardson and Robins (2014) argue that the price, supply and demand model cannot be modeled at equilibrium as an SCM without self-cycles. We conclude that it can be modeled by an SCM that contains self-cycles, with the corresponding graph depicted in Figure 3.6 (right).

A consequence of Theorem 3.4.11 is that if the SCM $\mathcal{M}_{\mathcal{R}}$ has no solutions, then the SDCM $\mathcal{R}$ has no equilibrating solutions. However, the converse does not hold in general, as the following example illustrates.

Example 3.4.14. *Let $\mathcal{R} = \langle \{1,2\}, \{3\}, \boldsymbol{\mathcal{X}}, \mathcal{E}, \boldsymbol{n}, \boldsymbol{f}, E \rangle$ be the steady SDCM with $\boldsymbol{\mathcal{X}} = \mathbb{R}^2$, $\mathcal{E} = \mathbb{R}$, $\boldsymbol{n} = (0,1)$, the dynamic causal mechanism $\boldsymbol{f}$ given by $f_1(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, e) = x_{2'}$ and*

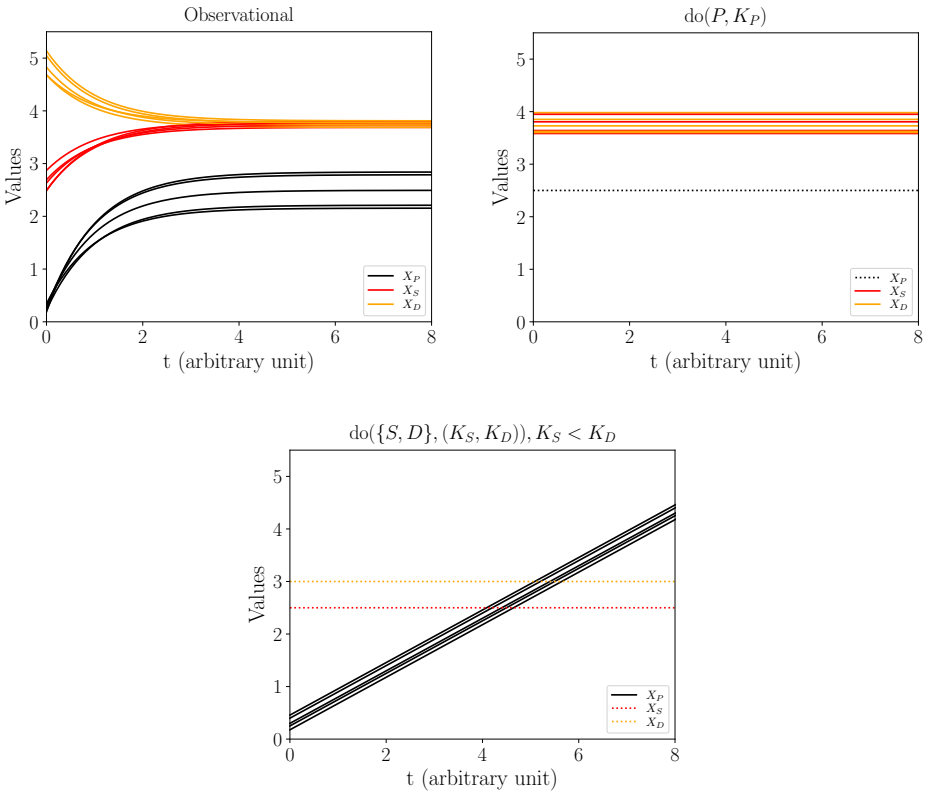


Figure 3.9: Simulation of solutions of the SDCM of the price, supply and demand model of Example 3.4.13 under different steady perfect interventions.

$f_2(\bar{\boldsymbol{x}}^{(\boldsymbol{n})}, e) = e$, and the exogenous process E given by $E_t = \sin(t^3)/t$. The dynamic structural equations associated to $\mathcal{R}$ are given by

$$X_1 = X_2', \quad X_2 = E\,.$$

This model can be equilibrated to the model $\mathcal{M}_{\mathcal{R}}$ with structural equations

$$X_1^* = 0\,, \quad X_2^* = E^*\,,$$

and exogenous variable $E^* = 0$. Although the SCM $\mathcal{M}_{\mathcal{R}}$ clearly has a solution, the SDCM $\mathcal{R}$ has no equilibrium states, since $X_1 = X_2' = E'$ is not a.s. convergent.

The following result shows that if the exogenous process is constant in time, this cannot happen.

Proposition 3.4.15. *Let $\mathcal{R}$ be a steady SDCM such that the exogenous process $\boldsymbol{E}$ is a random variable (i.e., $\boldsymbol{E}$ is constant in time). If the SDCM $\mathcal{R}$ has no equilibrating solution, then its equilibration $\mathcal{M}_{\mathcal{R}}$ has no solutions.*

3.4.4 *Graphs of the equilibrated SDCM*

In this subsection, we show how the equilibration operation acts on the (augmented) graph of the SDCM.

Proposition 3.4.16 (Graph of the equilibrated SDCM is a subgraph of the original mixed graph)**.** *Let $\mathcal{R}$ be a steady SDCM. The graph $\mathcal{G}(\mathcal{M}_{\mathcal{R}})$ of the equilibrated SDCM $\mathcal{M}_{\mathcal{R}}$ is the mixed graph obtained from the graph $\mathcal{G}(\mathcal{R})$ of $\mathcal{R}$ by removing the partition into clusters and removing the nodes $i^{(k_i)}$ for $i \in I$ and $k_i > 0$ together with their adjacent edges. An analogous statement holds for the augmented graph $\mathcal{G}^a(\mathcal{M}_{\mathcal{R}})$.*

The following example illustrates this for the equilibrated price, supply and demand model.

Example 3.4.17 (Price, supply and demand, continued)**.** *Consider the price, supply and demand model $\mathcal{R}$ of Example 3.3.15 for a very large λ, that is, for which the price adjusts very quickly to changes in supply and demand. This system can be approximated by the equilibrated price, supply and demand model $\mathcal{M}_{\mathcal{R}}$. The graph of this equilibrated model $\mathcal{M}_{\mathcal{R}}$ is a subgraph of the graph of the original model $\mathcal{R}$, as can be seen in Figure 3.6.*

3.4.5 *Equilibration commutes with intervention*

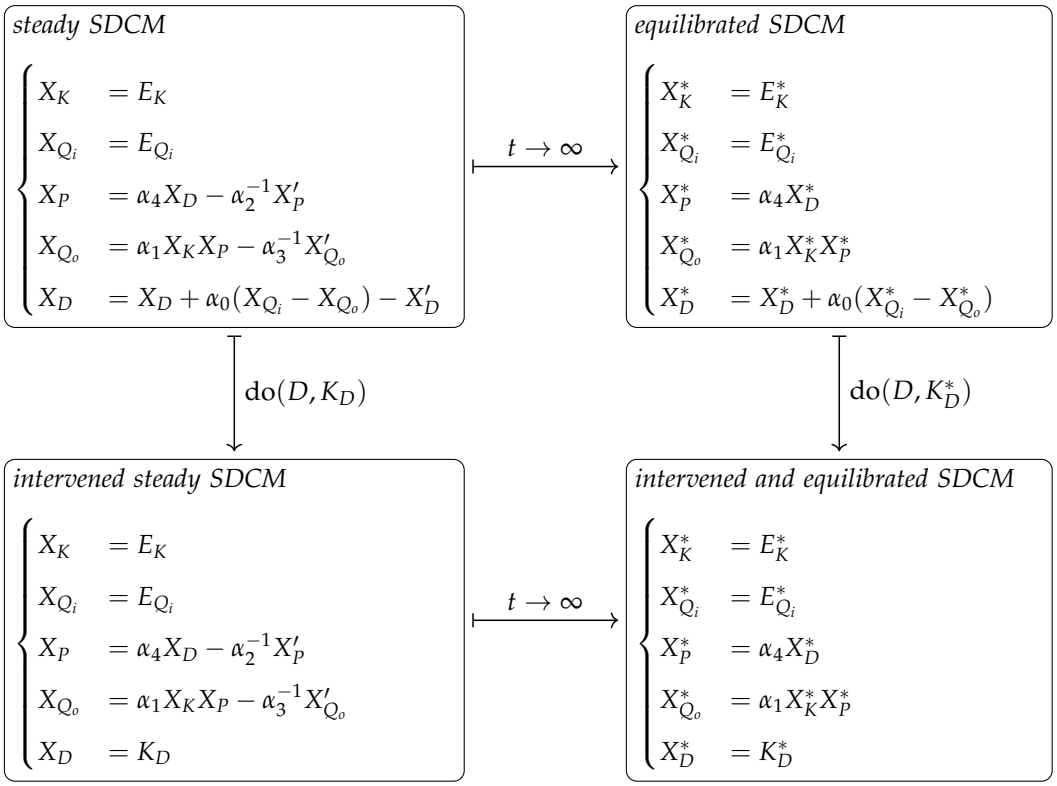
Theorem 3.4.11 states that the equilibrium states of a steady SDCM are solutions of the SCM to which the SDCM equilibrates. In the previous subsection, we showed, moreover, that the functional relationships between the endogenous processes that are encoded in the dynamic structural equations are preserved under equilibration. This leads to another important result: the equilibration operation preserves the causal semantics of the equilibrium states, as is illustrated in Figure 3.1 in Section 3.1.

Theorem 3.4.18. *Let $\mathcal{R}$ be a steady SDCM and let $\mathrm{do}(I, \boldsymbol{K}_I)$ be a steady stochastic perfect intervention for some subset $I \subseteq \mathcal{I}$ and stochastic process $\boldsymbol{K}_I$ that converges a.s. to a random variable $\boldsymbol{K}_I^*$. Then the steady stochastic perfect intervention commutes with equilibration, that is*

$$(\mathcal{M}_{\mathcal{R}})_{\mathrm{do}(I,\boldsymbol{K}_I^*)} = \mathcal{M}_{(\mathcal{R}_{\mathrm{do}(I,\boldsymbol{K}_I)})}\,.$$

This result allows us to perform causal reasoning on the equilibrium states of the SDCM by considering only the equilibrated model, as is illustrated in the following example.

Example 3.4.19 (Bathtub model, continued)**.** *In Example 3.3.30 we defined the SDCM for the bathtub model. The equilibrium states of this model can be described by the structural equations of the equilibrated model, as depicted in the top row of the following diagram.*



After equilibration, one can perform causal reasoning on the level of the equilibrated SDCM, without needing to resort to the original SDCM description. Indeed, we see in the above diagram that it doesn't matter whether we first perform the steady stochastic perfect intervention $\mathrm{do}(D, K_D)$, and then let the system equilibrate, or the other way around. The graphs of the SDCM, the equilibrated SDCM and their corresponding intervened models are depicted in Figure 3.7. Choosing different a.s. convergent processes for K_D yields different solution processes X_P and X_{Q_o} of the intervened SDCM, but the solution processes for X_{Q_i} and X_K stay unchanged. Similarly, the perfect intervention $\mathrm{do}(D, K_D^*)$ on the equilibrated SDCM yields different solutions X_P^* and $X_{Q_o}^*$ of the intervened SCM depending on the value of K_D^*, but does not change the solutions $X_{Q_i}^*$ and X_K^*. This behavior is also reflected in the graphs depicted in the bottom row of Figure 3.7.

Intuitively, one would indeed expect the chosen intervention value for the depth to have an effect on pressure and outflow (but not on inflow or drain size) at equilibrium. For example, one could (approximately) implement such a perfect intervention by adding a water level control device that constantly monitors the level and that can pump water in and out of the bathtub via a hose, regulating the depth at K_D at all times by using an optimal control feedback loop, independently of the exogenous processes E_K and E_{Q_i}. Indeed, the depth directly determines the pressure X_P exerted by the water in the bathtub at the drain, and the outflow rate X_{Q_o} is a direct consequence of that. Once the other processes in the system have equilibrated, the processes X_P and X_{Q_o} will also equilibrate to random variables that depend on K_D^*. The inflow $X_{Q_i}^*$ of water through the faucet no longer needs to be equal to the outflow $X_{Q_o}^*$ through the drain at equilibrium because water is also constantly added or removed via the hose by the water level control device in order to maintain the (eventually) constant depth K_D^*.[31]

This sheds some new light on the violation of the equilibration-manipulation commutability property (the "EMC-property") of Dash (2005), who shows the—at first sight contradictory—result that equilibration does not always commute with intervention. The paradox is resolved by noting that Dash (2005) defines a different notion of "equilibration", inspired by Iwasaki and Simon (1994), for which commutativity with perfect intervention indeed does not always hold. One can readily verify that the "equilibration" operation of Dash (2005) does not preserve the functional relationships between the endogenous processes that are encoded in the equations under the equilibration. Recently, Blom and Mooij (2021) showed that the "equilibration" operation of Dash (2005) maps an SDCM $\mathcal{R}$ to a Markov ordering graph that encodes the conditional independencies in solutions of $\mathcal{M}_{\mathcal{R}}$ instead of the functional relationships. In contrast, our equilibration operation, defined in Definition 3.4.9, preserves the functional relationships between the endogenous processes, since each dynamic structural equation equilibrates to a structural equation associated to the same endogenous process/variable. This is also reflected in Proposition 3.4.16 where we showed that the graph of the equilibrated SDCM is a subgraph of the mixed graph of the SDCM.

Theorem 3.4.11 and 3.4.18 together imply that our equilibration operation preserves the equilibrium states of a steady SDCM while also preserving the causal semantics. In particular, we do not require that all solutions of the steady SDCM have to equilibrate. As a consequence, the equilibrium states of the model may depend on the (consistent) initial conditions. This is in contrast to the work of Mooij, Janzing, and Schölkopf (2013), who assume that the equilibrium state of the dynamical system is unique and independent of the initial condition. This is a strong assumption that limits the applicability of the theory, since this does not allow for any stochasticity at equilibrium. Indeed, many random dynamical systems have multiple equilibrium states that depend on the chosen initial condition, as is illustrated in the following example.

31 At equilibrium, the total inflow of water through the faucet and the hose has to be equal to the total outflow through the drain and the hose.

Example 3.4.20 (Bathtub model, continued)**.** *Consider again the bathtub model $\mathcal{R}$ of Example 3.3.30. Figure 3.10 (top left) illustrates some numerical solutions of the dynamic SEs, with $\boldsymbol{\alpha} = (1, 1, 1, 1, 4/5)$, $E_K = 1/2, E_{Q_i} = 1$ and for randomly drawn consistent initial conditions $(0, \boldsymbol{X}_{[0]})$ of order $\mathbf{0}$. We see that the solutions equilibrate to the a.s. unique equilibrium state $(X^*_K, X^*_{Q_i}, X^*_P, X^*_{Q_o}, X^*_D) = (1/2, 1, 2, 5/2, 1)$ corresponding to the solution of the equilibrated SDCM $\mathcal{M}_\mathcal{R}$. If we now perform the perfect intervention $\mathrm{do}(Q_o, K_{Q_o})$ on the system $\mathcal{R}$, where we force the water outflow X_{Q_o} to be equal to the water inflow X_{Q_i} at all time, that is, $K_{Q_o} = E_{Q_i}$, then this does not give an a.s. unique equilibrium state, but the equilibrium state that is obtained depends on the initial condition, as can be seen in Figure 3.10 (center left). Indeed, the depth X^*_D at equilibrium will equal the initial depth $X_{[0],D}$ at $t_0 = 0$, if the inflow X_{Q_i} equals the outflow X_{Q_o}. This example cannot be treated with the theory of (Mooij, Janzing, and Schölkopf, 2013), which assumes that the equilibrium state is unique and does not depend on the initial condition. However, if instead we perform the perfect intervention $\mathrm{do}(Q_o, K_{Q_o})$ on $\mathcal{R}$ where $K_{Q_o} < E_{Q_i}$, then the depth X_D will not reach equilibrium, but will increase indefinitely, since the rate of water flowing into the bathtub is larger than the outflow rate. This is illustrated in Figure 3.10 (center right). This is also reflected in the equilibrated SDCM $\mathcal{M}_\mathcal{R}$, which does not have any solution after the corresponding perfect intervention $\mathrm{do}(Q_o, K_{Q_o})$.*

Similar behavior is observed for the equilibrium states of the price, supply and demand model $\mathcal{R}$ of Example 3.4.13. For example, the model $\mathcal{R}$ will reach market equilibrium if one holds the price fixed at all times by the perfect intervention $\mathrm{do}(P, K_P)$, but will not reach equilibrium if the supply and demand are fixed at all times by the perfect intervention $\mathrm{do}(\{S, D\}, (K_S, K_D))$ for which $K_S < K_D$ (see Figure 3.9 top right and bottom respectively). In all the cases depicted in Figure 3.9 we see a dependence of the equilibrium states on the initial condition.

In summary, the equilibration of a steady SDCM to an SCM generalizes the work of (Mooij, Janzing, and Schölkopf, 2013) in three directions: (i) the deterministic setting is replaced with a more general stochastic setting, (ii) the dynamic structural equations can be of arbitrary order (including zeroth-order), rather than only first-order, which prevents complications with the causal interpretation (see, for example, Example 3.3.11), and (iii) the equilibrium state is allowed to depend on initial conditions. Together, this substantially extends the applicability of the theory.

3.4.6 *Realizing a given SCM as a stable SDCM*

Although each steady SDCM equilibrates to an SCM, not all solutions of the SDCM need to equilibrate to solutions of the corresponding SCM (see, for example, Example 3.4.12). In this subsection, we address the inverse problem of finding steady SDCMs with non-trivial dynamics for which all solutions equilibrate to solutions of a specified SCM. This can be thought of as realizing the given SCM as a "stable" SDCM. In Proposition 3.4.4 we provided certain conditions under which all the solutions of a linear SDCM equilibrate. Based on this result and some results in the linear systems theory literature, we show that for a certain class of SCMs one can construct a first-order SDCM such that *all* its solutions equilibrate to the

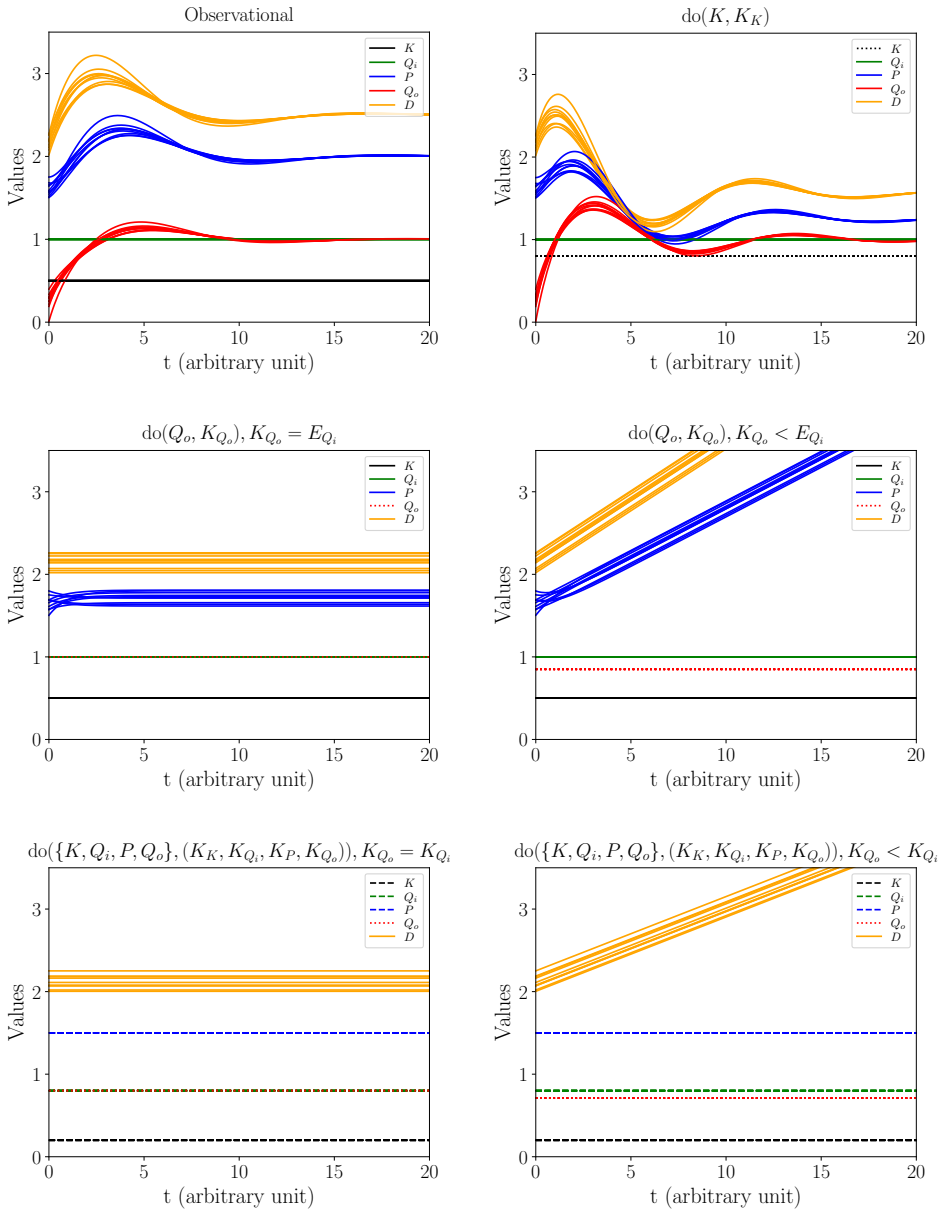


Figure 3.10: Simulation of solutions of the SDCM of the bathtub model of Example 3.4.20 and 3.4.25 under different steady perfect interventions.

solutions of the SCM. Moreover, we show that under certain stronger conditions, the SDCM can be chosen such that its solutions still equilibrate to the solutions of the intervened SCM after any constant stochastic perfect intervention. Hence, the constructed SDCM realizes the causal semantics of the SCM at equilibrium.

First, we observe that one cannot uniquely recover an SDCM from its equilibration in general.

Example 3.4.21. *Consider the linear SDCM $\mathcal{R}$ with dynamic SE given by*

$$\boldsymbol{X} = B\boldsymbol{X} - \boldsymbol{X}' + \Gamma\boldsymbol{E}\,,$$

where the matrix $A := \mathbb{I} - B$ is invertible and the exogenous process $\boldsymbol{E}$ is a random variable. Consider another SDCM $\tilde{\mathcal{R}}$ which differs only in its dynamic causal mechanism, and has the dynamic SE

$$\boldsymbol{X} = B\boldsymbol{X} - \Lambda\boldsymbol{X}' + \Gamma\boldsymbol{E}\,, \tag{3.6}$$

where Λ is some invertible diagonal matrix. The equilibrated SDCMs $\mathcal{M}_{\mathcal{R}}$ and $\mathcal{M}_{\tilde{\mathcal{R}}}$ coincide, and have structural equations of the form

$$\boldsymbol{X}^* = B\boldsymbol{X}^* + \Gamma\boldsymbol{E}^*\,. \tag{3.7}$$

Hence, the equilibrium states $\boldsymbol{X}^*$ of the SDCMs $\mathcal{R}$ and $\tilde{\mathcal{R}}$ are indistinguishable, since both have to satisfy $\boldsymbol{X}^* = A^{-1}\Gamma\boldsymbol{E}^*$ a.s.. Furthermore, if the matrix $\mathbb{I}_I - B_{II}$ is invertible for some subset $I \subseteq \mathcal{I}$, then also for $J := \mathcal{I} \setminus I$ the intervened equilibrium states of $\mathcal{R}_{\mathrm{do}(J,\boldsymbol{K}_J)}$ and $\tilde{\mathcal{R}}_{\mathrm{do}(J,\boldsymbol{K}_J)}$ are indistinguishable for any sufficiently smooth steady stochastic intervention $\mathrm{do}(J,\boldsymbol{K}_J)$.

Although the equilibrated SDCM in Example 3.4.21 describes the possible equilibrium state of both SDCMs, it is not necessarily guaranteed that the solutions of both SDCMs equilibrate. One might hope that for any given linear SCM of the form (3.7), one can always find an invertible diagonal matrix Λ such that one can construct a steady SDCM of the form (3.6) for which all solutions of the SDCM equilibrate to the (a.s. unique) solution of the SCM (see Proposition 3.4.4). Such a "stabilization matrix" Λ does not always exist. A sufficient condition for its existence was given in (Fisher and Fuller, 1958; Fisher, 1972), leading to the following result.

Corollary 3.4.22. *Let $\mathcal{M}$ be a linear SCM with structural equations*

$$\boldsymbol{X} = B\boldsymbol{X} + \Gamma\boldsymbol{E}\,,$$

where $\mathcal{I} = \{1,\dots,d\}$, $\mathcal{J} = \{1,\dots,e\}$, $B \in \mathbb{R}^{d\times d}$, $\Gamma \in \mathbb{R}^{d\times e}$, and with $\boldsymbol{E}$ a random variable. Write $A := \mathbb{I} - B$. For a diagonal matrix $\Lambda \in \mathbb{R}^{d\times d}$, consider the linear SDCM $\mathcal{R}_{\mathcal{M},\Lambda}$ with dynamic SE of the form

$$\boldsymbol{X} = B\boldsymbol{X} - \Lambda\boldsymbol{X}' + \Gamma\boldsymbol{E}.$$

If there exists a sequence of matrices $M_d, M_{d-1}, \ldots, M_1$ with $M_d = A$ such that for $k = 2, \ldots, d$ each M_{k-1} is a principal $(k-1) \times (k-1)$ submatrix of M_k, with $\det M_k \neq 0$ for all $k = 1, \ldots, d$, then there exists a diagonal stabilization matrix $\Lambda \in \mathbb{R}^{d \times d}$ such that the linear SDCM $\mathcal{R}_{\mathcal{M},\Lambda}$ has the properties that (i) its equilibrated SDCM is $\mathcal{M}_{\mathcal{R}_{\mathcal{M},\Lambda}} = \mathcal{M}$, and (ii) all its solutions equilibrate to an a.s. unique equilibrium state that satisfies the structural equations of the SCM $\mathcal{M}$, independent of the initial condition.

While this sufficient condition guarantees the existence of a stabilization matrix Λ such that the *observational* equilibrium distribution of the SCM is recovered as the distribution of the equilibrium state of the SDCM, it does not guarantee that after a stochastic perfect intervention on the SDCM, all solutions will equilibrate to an (a.s. unique) equilibrium solution of the corresponding intervened SCM. Indeed, a certain Λ that stabilizes the dynamics in the absence of the intervention may no longer stabilize the dynamics after the intervention has been carried out. Can we, under some conditions, find a single Λ that will stabilize the dynamics after *any* stochastic perfect intervention? The answer is affirmative, as was shown by Locatelli and Schiavoni (2012) who provide a necessary and sufficient condition for the existence of an invertible diagonal stabilization matrix Λ that simultaneously stabilizes all subsystems.[32] This leads to the following result on how one can "realize" a given linear SCM as a stable linear first-order SDCM.

Corollary 3.4.23. *Let $\mathcal{M}$ be a linear SCM with structural equations*

$$\boldsymbol{X} = B\boldsymbol{X} + \Gamma \boldsymbol{E},$$

where $\mathcal{I} = \{1, \ldots, d\}$, $\mathcal{J} = \{1, \ldots, e\}$, $B \in \mathbb{R}^{d \times d}$, $\Gamma \in \mathbb{R}^{d \times e}$, and with $\boldsymbol{E}$ a random variable. Write $A := \mathbb{I} - B$. For a diagonal matrix $\Lambda \in \mathbb{R}^{d \times d}$, consider the linear SDCM $\mathcal{R}_{\mathcal{M},\Lambda}$ with dynamic SE of the form

$$\boldsymbol{X} = B\boldsymbol{X} - \Lambda \boldsymbol{X}' + \Gamma \boldsymbol{E}.$$

If

$$\det(A_{II}) \det(\operatorname{diag}(A_{II})) > 0 \qquad \forall I \subseteq \mathcal{I}, \tag{3.8}$$

then there exists an invertible diagonal stabilization matrix $\Lambda \in \mathbb{R}^{d \times d}$ such that the linear SDCM $\mathcal{R}_{\mathcal{M},\Lambda}$ has the properties that (i) its equilibrated SDCM is $\mathcal{M}_{\mathcal{R}_{\mathcal{M},\Lambda}} = \mathcal{M}$, and (ii) under every stochastic perfect intervention $\operatorname{do}(J, \boldsymbol{K}_J)$ with the exogenous process $\boldsymbol{K}_J$ constant in time, all solutions of $(\mathcal{R}_{\mathcal{M},\Lambda})_{\operatorname{do}(J,\boldsymbol{K}_J)}$ equilibrate to an a.s. unique equilibrium state that is the unique solution of the SCM $\mathcal{M}_{\operatorname{do}(J,\boldsymbol{K}_J^*)}$, independent of the initial condition.

Condition (3.8) implies that the matrices $\mathbb{I}_I - B_{II}$ are invertible for every subset $I \subseteq \mathcal{I}$. Such linear SCMs are special cases of the class of *simple* SCMs (see Section 2.8).

32 Locatelli and Schiavoni (2012) consider an extension of the stabilization problem studied by Fisher and Fuller (1958). Whereas Fisher and Fuller (1958) consider the problem of finding a diagonal matrix $\Lambda \in \mathbb{R}^{d \times d}$ for a matrix $A \in \mathbb{R}^{d \times d}$ such that the matrix ΛA is Hurwitz (for which they provide a sufficient condition), Locatelli and Schiavoni (2012) consider the case where *all* the principal submatrices of ΛA should be Hurwitz, and provide a condition that is both sufficient and necessary, as well as a construction of such a stabilization matrix Λ.

Simple SCMs have the convenient property that their solutions are a.s. unique after any stochastic perfect intervention. We conclude that for the subclass of simple linear SCMs that satisfy condition (3.8), we can construct a linear first-order SDCM whose causal semantics at equilibrium "realizes" that described by the SCM. We speculate that this result can be extended to higher-order and nonlinear systems, but we will not pursue these questions here.

Example 3.4.24. *We show that the equilibrated SDCM of Example 3.4.12 (see also Example 3.3.5), modeling the equilibrium states of a damped coupled harmonic oscillator, satisfies condition (3.8). Indeed, taking $\mathcal{I} = \{1, \dots, d\}$, the matrix B of this linear SCM is tridiagonal, given as*

$$B = \begin{bmatrix} 0 & \frac{\kappa_1}{\kappa_0+\kappa_1} & & & \\ \frac{\kappa_1}{\kappa_1+\kappa_2} & 0 & \frac{\kappa_2}{\kappa_1+\kappa_2} & & \\ & \frac{\kappa_2}{\kappa_2+\kappa_3} & 0 & \ddots & \\ & & \ddots & \ddots & \frac{\kappa_{d-1}}{\kappa_{d-2}+\kappa_{d-1}} \\ & & & \frac{\kappa_{d-1}}{\kappa_{d-1}+\kappa_d} & 0 \end{bmatrix},$$

where $\kappa_0 = \kappa_d = 0$. Hence $A = \mathbb{I} - B = DC$ with diagonal

$$D = \begin{bmatrix} \frac{1}{\kappa_0+\kappa_1} & & & & \\ & \frac{1}{\kappa_1+\kappa_2} & & & \\ & & \frac{1}{\kappa_2+\kappa_3} & & \\ & & & \ddots & \\ & & & & \frac{1}{\kappa_{d-1}+\kappa_d} \end{bmatrix}$$

and tridiagonal

$$C = \begin{bmatrix} \kappa_0+\kappa_1 & -\kappa_1 & & & \\ -\kappa_1 & \kappa_1+\kappa_2 & -\kappa_2 & & \\ & -\kappa_2 & \kappa_2+\kappa_3 & \ddots & \\ & & \ddots & \ddots & -\kappa_{d-1} \\ & & & -\kappa_{d-1} & \kappa_{d-1}+\kappa_d \end{bmatrix}.$$

The determinants of D and C can be expressed in closed form as

$$\det D = \prod_{i=1}^{d} \frac{1}{\kappa_{i-1}+\kappa_i}, \qquad \det C = \sum_{i=0}^{d} \prod_{\substack{j=0 \\ j\neq i}}^{d} \kappa_j.$$

Hence, since $\kappa_i > 0$ for $i = 1, \dots, d-1$, $\det A = (\det C)(\det D) > 0$. Also, we clearly have $\det \operatorname{diag}(A) > 0$. Hence, condition (3.8) holds for $I = \mathcal{I}$. A similar calculation (and

exploiting the block structure of the principal submatrices) shows that condition (3.8) holds for all $I \subseteq \mathcal{I}$.

Remarkably, we can thus apply Corollary 3.4.23 to the damped harmonic oscillator SCM to obtain a realization of this causal equilibrium model as a *first-order* linear SDCM (remember that the original SDCM is a second-order linear SDCM).

3.4.7 *Causal interpretation of the graph of the equilibrated SDCM*

While the graph of an acyclic SCM has a straightforward causal interpretation, this need not be the case for general SCMs with cycles (see Chapter 2).[33] While an acyclic SCM induces a unique "observational" distribution, cyclic SCMs may induce none, one or several different observational distributions. Similarly, after performing a perfect intervention on some of the variables, a cyclic SCM may induce none, one or several different corresponding interventional distributions. In general, one has to be careful in how to causally interpret the graph of an SCM if cycles are present; in particular, this caveat holds for SCMs that are obtained as the equilibration of an SDCM. First, not all directed edges and directed paths in the graph can easily be identified from differences in interventional distributions in case cycles are present (see Section 2.7). Second, if cycles are present, "nonancestral" effects may exist (see also Section 2.7), that is, an intervention on a variable may change the distribution of some of its nondescendants in the graph. In this subsection, we show how these subtleties and counterintuitive nonancestral effects in cyclic equilibrated SDCMs can be explained in terms of properties of the underlying SDCM.

In general, the presence or absence of a directed edge or path in the graph of an SCM $\mathcal{M}$ cannot always be identified from the observational and interventional distributions. In the cyclic setting, the following sufficient condition can be used to identify such directed edges or paths between nodes i and j (see Proposition 2.7.1 for the exact formulation).

- A *direct causal effect* of i on j can be identified, that is, there exists a $i \to j \in \mathcal{G}(\mathcal{M})$, if (i) the structural equation of j can be solved a.s. uniquely for X_j in terms of the other variables that appear in the equation, and (ii) there exist values $\boldsymbol{K}_I \in \boldsymbol{\mathcal{X}}_I$ and $K_i \neq \tilde{K}_i \in \mathcal{X}_i$, where $I = \mathcal{I} \setminus \{i, j\}$, and a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that the following probabilities are uniquely defined and do not coincide:

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

- An *indirect causal effect* of i on j can be identified, that is, there exists a directed path $i \to \cdots \to j$ in $\mathcal{G}(\mathcal{M})$, if (i) the structural equations of the ancestors of j in $\mathcal{G}(\mathcal{M})_{\setminus i}$ (that is, the graph $\mathcal{G}(\mathcal{M})$ where we removed the node i and its adjacent edges) can be solved a.s. uniquely for their associated variables in

33 The straightforward causal interpretation of acyclic SCMs actually extends to a much more general class of possibly cyclic SCMs called *simple* SCMs (see Section 2.8).

directed path to

from	K	Q_i	P	Q_o	D
K	-	×	✓	✓	✓
Q_i	×	-	✓	✓	✓
P	×	×	-	✓	✓
Q_o	×	×	✓	-	✓
D	×	×	✓	✓	✓

directed edge to

from	K	Q_i	P	Q_o	D
K	-	×	×	✓	×
Q_i	×	-	×	×	✓
P	×	×	-	✓	×
Q_o	×	×	×	-	✓
D	×	×	✓	×	✓

according to $\mathcal{G}(\mathcal{M}_\mathcal{R})$ (see Figure 3.7 top right)

indirect effect

cause	K	Q_i	P	Q_o	D
K	-	?	✓	?	✓
Q_i	?	-	✓	✓	✓
P	??	??	-	??	??
Q_o	??	??	??	-	??
D	?	?	✓	✓	-

direct effect

cause	K	Q_i	P	Q_o	D
K	-	?	?	✓	??
Q_i	?	-	?	?	??
P	?	?	-	✓	??
Q_o	?	?	?	-	??
D	?	?	✓	?	-

Identifiable from the observational and interventional distributions with Prop. 2.7.1

Table 3.1: The directed paths/edges (top tables) of the equilibrated bathtub model $\mathcal{M}_\mathcal{R}$ and the (in)direct causal effects that can be identified by Proposition 2.7.1 (bottom tables) are denoted by a "✓". Those that cannot be identified are denoted by the question marks "?" and "??". A single question mark "?" denotes that condition (i) is satisfied, but not condition (ii), while a double question mark "??" denotes that condition (i) is not satisfied.

terms of the other variables that appear in these equations, and (ii) there exist values $K_i \neq \tilde{K}_i \in \mathcal{X}_i$ and a measurable set $\mathcal{B}_j \subseteq \mathcal{X}_j$ such that the following probabilities are uniquely defined and do not coincide:

$$\mathbb{P}_{\mathcal{M}_{\mathrm{do}(i,K_i)}}(X_j \in \mathcal{B}_j) \neq \mathbb{P}_{\mathcal{M}_{\mathrm{do}(i,\tilde{K}_i)}}(X_j \in \mathcal{B}_j)\,.$$

In the following example, we illustrate how we can interpret the directed edges and paths of the equilibrated bathtub model that cannot be identified by this sufficient condition from an SDCM perspective.

Example 3.4.25 (Bathtub model, continued)**.** *Consider again the bathtub model $\mathcal{R}$ of Example 3.3.30. We simulated some numerical solutions, with parameters as given in Example 3.4.20, shown in Figure 3.10 (top left). In Table 3.1 (bottom left) one can read off all the indirect causal effects that can be identified by comparing different interventional distributions from the equilibrated model $\mathcal{M}_\mathcal{R}$ with the help of Proposition 2.7.1. The indirect causal effects of P and Q_o cannot be identified by comparing interventional distributions, since the intervened equilibrated models $(\mathcal{M}_\mathcal{R})_{\mathrm{do}(P,K_P)}$ and $(\mathcal{M}_\mathcal{R})_{\mathrm{do}(Q_o,K_{Q_o})}$ do not have a solution (except for one special choice of K_P respectively K_{Q_o}), and hence condition (i) is not satisfied. This was already illustrated for the perfect intervention $\mathrm{do}(Q_o, K_{Q_o})$ in Figure 3.10 (center left/right) of Example 3.4.20.*

The direct causal effects that can be identified from $\mathcal{M}_\mathcal{R}$ are given in Table 3.1 (bottom right). The direct causes of D cannot be identified due to the self-cycle at D, which means that condition (i) is not satisfied, that is, the structural equation of D cannot be a.s. uniquely solved for D in terms of the other variables. Indeed, the depth D will not equilibrate, but

will increase indefinitely, if the rate of water into the bathtub is larger than the outflow rate, that is, $K_{Q_o} < K_{Q_i}$ (see Figure 3.10 bottom right). On the other hand, it will reach an equilibrium state only if the rate of water into and out of the bathtub are equal, that is, $K_{Q_o} = K_{Q_i}$. In this case, the depth D will remain constant over all times, as illustrated in Figure 3.10 (bottom left).

The directed path from K to Q_o in the graph of the equilibrated model $\mathcal{M}_{\mathcal{R}}$ cannot be straightforwardly identified as an indirect causal effect at equilibrium, because the equilibrium distribution of Q_o does not change due to perfect interventions on K (this corresponds to the single question mark in the Table 3.1, bottom left), as explicit calculations reveal. However, at some finite time point one does observe changes in the distribution of Q_o when performing perfect interventions on K (Figure 3.10 (top)). Together, this implies that this system is capable of perfect adaptation (Blom and Mooij, 2021). Interestingly, the direct edge $K \to Q_o$ in the graph of the equilibrated model $\mathcal{M}_{\mathcal{R}}$ can be identified by changes in the equilibrium distribution of Q_o under perfect interventions on K, Q_i, P, D (which then also implies that there is a directed path from K to Q_o in the graph of the equilibrated model).

In particular, this example illustrates that one can run into several problems when one attempts to identify directed edges and paths of the graph of the SCM from the differences in equilibrium distributions under interventions on the SDCM:

- if the intervened SCM has no solutions, then the descendants of the intervention targets cannot be easily identified;
- if the graph of the SCM has a self-cycle at some variable, then the parents of that variable cannot be easily identified;
- if the equilibrium distribution of some descendants of the intervention target variable remain insensitive to the intervention (for example, when the dynamical system exhibits perfect adaptation (Blom and Mooij, 2021)), these descendants cannot be easily identified.

In Example 3.4.25, the identified indirect causal relationships are a subset of the ancestral relationships. This can be seen from observing that each "✓" in Table 3.1 (bottom left) has a corresponding "✓" in Table 3.1 (top left). In other words, performing a perfect intervention on a variable can only change the distribution of its descendants in the graph. In general, however, it can happen that an intervention on a nonancestor of a variable can change the distribution of that variable (see Section 2.7). This counterintuitive behavior of "nonancestral" effects in an equilibrated SDCM can be explained by the dependence of the equilibrium states on the initial conditions in combination with the fact that not each initial condition corresponds to an equilibrating solution. The following example illustrates this.

Example 3.4.26 (Selection bias leading to nonancestral effects in an equilibrated SDCM)**.** *Consider the SDCM $\mathcal{R}$ with dynamic structural equations given by*

$$\begin{cases} X_1 = X_1 - X_1' + 2X_2 - X_3 \\ X_2 = X_2 - X_2' \\ X_3 = E\,, \end{cases}$$

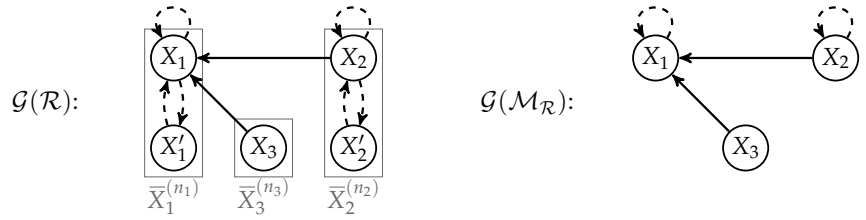


Figure 3.11: Graphs of the SDCM $\mathcal{R}$ (left) and the corresponding equilibrated model $\mathcal{M}_\mathcal{R}$ (right) of Example 3.4.26.

with order tuple $\boldsymbol{n} = (1,1,0)$ *and* E *some constant in* $\mathbb{R}$. *Denote* $I = \{1,2\}$ *and note that* $\mathcal{R}$ *satisfies Assumption 1-*$(I \subseteq \mathcal{I})$. *The equilibrated model* $\mathcal{M}_\mathcal{R}$ *is given by*

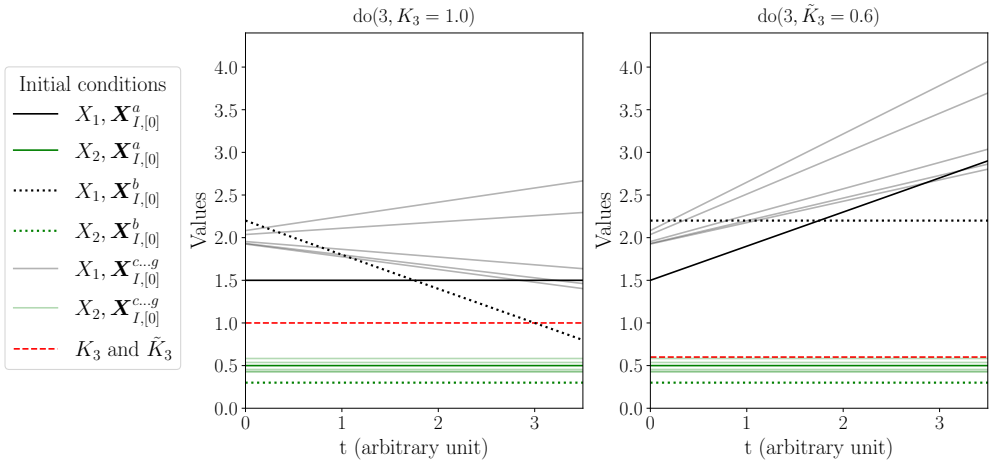
$$\begin{cases} X_1^* = X_1^* + 2X_2^* - X_3^* \\ X_2^* = X_2^* \\ X_3^* = E\,. \end{cases}$$

The graphs of $\mathcal{R}$ *and* $\mathcal{M}_\mathcal{R}$ *are depicted in Figure 3.11. First observe that the induced equilibrium distribution of* X_2^* *differs for two constant perfect interventions* $\mathrm{do}(3, K_3)$ *and* $\mathrm{do}(3, \tilde{K}_3)$ *with* $K_3 \neq \tilde{K}_3$, *since the equilibrium state has to satisfy* $X_2^* = X_3^*/2$ *a.s.. However, there is no directed path from the variable* X_3 *to the variable* X_2 *in the graph of the SCM* $\mathcal{M}_\mathcal{R}$. *This counterintuitive behavior can be explained by taking the initial conditions of the solutions of the SDCM into account, as we shall now explain.*

In Figure 3.12, we plot the solutions of the SDCM $\mathcal{R}$ *for different partial initial conditions* $(t_0, \boldsymbol{X}_{I,[0]}^i)$ *at* $t_0 = 0$ *(for* $i = a, b, \ldots, g$*) under two steady perfect interventions, namely* $\mathrm{do}(3, K_3 = 1.0)$ *and* $\mathrm{do}(3, \tilde{K}_3 = 0.6)$. *For illustration purposes, we consider here only non-random initial conditions, because we can then identify the initial conditions with "individual" solutions, as depicted in Figure 3.12 (note that Corollary 3.3.28 applies). Observe that the set of partial initial conditions that correspond to equilibrating solutions differs for the two interventions. For the intervened model* $\mathcal{R}_{\mathrm{do}(3,K_3)}$, *the only solution that equilibrates is the one with initial condition* $(t_0, \boldsymbol{X}_{I,[0]}^a)$ *(denoted by the dark solid lines Figure 3.12 (top left)), whereas for the intervened model* $\mathcal{R}_{\mathrm{do}(3,\tilde{K}_3)}$ *the only solution that equilibrates is the one with initial condition* $(t_0, \boldsymbol{X}_{I,[0]}^b)$ *(denoted by the dark dotted lines in Figure 3.12 (top right)). This explains the counterintuitive behavior of nonancestral effects in the equilibrium SCM: The chosen value for* X_3 *affects which solutions will equilibrate, and thereby affects the equilibrium distribution of* X_2.

Note that at any finite point in time, these "nonancestral" effects do not occur; indeed, Figure 3.12 shows that the distribution of X_1 *differs for the two interventions at finite time, while that of* X_2 *remains unaffected.*

This example shows that the nonancestral effects in an equilibrated SDCM can be explained by the dependence of the equilibrium states on the initial conditions, in combination with the fact that not each initial condition corresponds to an equilibrating solution. Another way to think about this is as selection bias due to the assumption that the system has reached equilibrium. An intervention targeting



	$(X^i_{I,[0]})_1$	$(X^i_{I,[0]})_2$
$i = a$	1.5	0.5
$i = b$	2.2	0.3
$i = c \ldots g$	random	random

Figure 3.12: Simulation of solutions of the SDCM of Example 3.4.26 under different steady perfect interventions on X_3 (top left and right). The simulations in the top left and right plots are performed under the same set of initial conditions, summarized in the bottom table, but under different interventions.

a certain variable may change the set of equilibrating initial conditions of the system, and it can even change initial conditions for non-ancestors of the intervention target. By only considering these equilibrating initial conditions, this may appear as a causal effect of a variable on some of its non-ancestors at equilibrium. When seen from this perspective, these "causal effects" can be considered to be spurious as they do not appear on an "individual level", that is, for individual trajectories (at finite time t), but only appear on a "population level" when selecting on some later event (namely, the system being at equilibrium). One can indeed think of this as selection bias due to equilibration.

3.5 DISCUSSION

Dynamical models consisting of (ordinary or random) differential equations are widely applied in science and engineering to model the dynamics of systems that are composed of several components. These differential equations by themselves do not have a clearcut causal interpretation. Although they may implicitly explain a particular phenomenon in terms of its causes, the causal semantics of the constituent components are generally not explicitly defined without additional assumptions.

In this work, we introduced structural dynamical causal models that formally encode causal semantics of stochastic processes by means of a structured set of random differential equations. SDCMs can be seen as stochastic-process versions of

structural causal models, where the random variables are replaced by stochastic processes and their derivatives. By viewing the (higher-order) derivatives $X_i^{(k_i)}$ to be aspects of the process X_i we arrive at a natural causal interpretation, where it is not necessary to even consider questions like "does position cause velocity, or does velocity cause position, or both?".

For steady SDCMs (for which the explicit time-dependence of the dynamics vanishes as $t \to \infty$) we introduced an equilibration operation that equilibrates the dynamic causal mechanism of each component separately. This led to the important result that intervention and equilibration commute, thus connecting the causal semantics at equilibrium with the causal semantics of the dynamics. It generalizes the analogous result of Mooij, Janzing, and Schölkopf (2013) in three directions: (i) we replaced the deterministic setting with a more general stochastic setting, which allows us to address both cycles and confounders, (ii) we allowed the order of the dynamic structural equations to be arbitrary, including zeroth-order, rather than only allowing for first-order differential equations, and (iii) we have dropped the strong assumption that the dynamical model needs to have a single globally attractive equilibrium state. This allows us to study the causal semantics of the equilibrium states of a plethora of dynamical systems subject to time-varying random disturbances encountered in science and engineering within the framework of structural causal models.

Our commutation result may appear to be at odds with the possible "violation of the equilibration-manipulation commutability property" pointed out by Dash (2005). Under our notion of equilibration—contrary to that of Dash—each dynamic structural equation of the SDCM becomes a structural equation of the SCM. This one-to-one correspondence between the equations leads to the preservation of the causal semantics under equilibration. We can reinterpret the phenomenon that Dash observed as the fact that the equilibrium distributions of certain dynamical systems (for example ones that exhibit perfect adaptation) are not faithful to the graph of the equilibrated SCM (Blom and Mooij, 2021), in the sense that they can contain conditional independencies not explained by this graph. For dynamical systems exhibiting perfect adaptation, these faithfulness violations are due to the structure of the dynamics, rather than "accidental" parameter cancellations. This has serious repercussions for attempts at inferring the causal structure from (conditional independencies in) equilibrium data (Blom and Mooij, 2021). Thus, in a different way we arrive at the same conclusion as Dash obtained.

In comparison with the causal constraints models of Blom, Bongers, and Mooij (2019), our modeling framework is more "agnostic" as we decided not to incorporate the initial conditions into the model.[34] This allowed us to causally model *all* the equilibrium states of a steady SDCM with a single SCM. However, that single SCM may not provide a *complete* description of the causal semantics at equilibrium (Blom, Bongers, and Mooij, 2019). This is indeed a modeling tradeoff: the simpler structure of SCMs compared to that of causal constraints models can come at the cost of a less complete description of the equilibrium behavior of certain dynamical systems.

34 This is analogous to the difference between an ODE and an initial-value problem.

On the other hand, the connection between the structure of the SCM and that of the underlying SDCM is straightforward, whereas it is not well understood at present how one can easily derive a concise yet complete representation of an equilibrated SDCM (and a corresponding initial condition) as a causal constraints model.

However, allowing for multiple (or no) solutions also comes at a cost: the causal interpretation of the SCM is more subtle than that of acyclic (or more generally, simple) SCMs, and in particular, does not straightforwardly relate to properties of its graph. We illustrated for the bathtub model how one can causally interpret the directed edges and paths of the graph of the SCM that models the equilibrium states of the underlying SDCM. We saw that one may run into several problems when attempting to identify aspects of the SCM graph from comparing differences in equilibrium distributions after intervening on some of the variables:

- if the intervened SCM has no solution (which may happen if the intervened SDCM does not converge to a finite equilibrium state, but instead diverges to infinity, or reaches a periodic limit cycle, for example), descendants of the intervention targets cannot be easily identified;
- if the SCM graph has a self-cycle at some variable (which may happen if the causal mechanism for that variable does not equilibrate for certain values of its parents), then the parents of that variable cannot be easily identified;
- if the equilibrium distributions of some descendants of the intervention target variable remain insensitive to the intervention (which may happen in dynamical systems exhibiting perfect adaptation), these descendants cannot be easily identified.

Even worse, the equilibrium SCM may entail distribution changes under interventions that appear to be of a causal nature, while no corresponding causal relations are present in the dynamics (and therefore, no corresponding ancestral relations are present in the SCM graph), as we pointed out in Example 3.4.26. These counterintuitive "nonancestral causal effects" can be understood as arising from the implicit selection bias due to conditioning on the system having reached an equilibrium state. Indeed, the solutions of the equilibrium SCM correspond to those solutions of the SDCM that have equilibrated, while the non-equilibrating solutions of the SDCM are ignored. In other words, the SCM provides the "population-level" causal semantics of the population of equilibrating SDCM solutions (at $t = \infty$), which can deviate from the "individual-level" causal semantics of (possibly non-equilibrating) SDCM solutions (at finite t). The phenomenon that population-level causality may differ from individual-level causality due to post-intervention selection bias is well-known in other contexts. For example, a car mechanic who only observes cars that don't start may conclude that replacing the battery causes start engines to fail. While this appears as a genuine causal effect on the population level, it would be foolish to conclude that this causal effect also pertains to individual cars. Intuitively, one might prefer to interpret such phenomena as not representing "truely causal" relations. On the other hand, if one is only interested in the effects

of interventions on a population level, there seems to be no harm in considering these distribution changes as causal. Thus, as long as one is explicit whether one refers to population-level or individual-level causality, both notions of causality can meaningfully co-exist. The important take-away, from our point of view, is that focussing on equilibrated systems may lead to selection bias.

As a side note, Example 3.4.26 also shows that SCMs may not fully capture such population-level causal relations graphically. We note that the recently proposed framework of Blom, Diepen, and Mooij (2021) is better suited in general to read off such population-level causal effects graphically from the structure of the equilibrium equations, under certain "local" solvability assumptions on these equations (rather than having to study global solutions of intervened equilibrium equations, as we did here).

Apart from these subtleties regarding their causal semantics, SCMs with cycles bring about several other challenges in general. For example, they generally do not have a Markov property, and the class of cyclic SCMs is not closed under marginalization. The subclass consisting of simple SCMs (see Section 2.8) allows for cycles, but simple SCMs share many of the convenient properties of acyclic SCMs. Hence, these convenient properties are directly applicable to the equilibrium states of those steady SDCMs that equilibrate to a simple SCM. This enables one to study the equilibrium states of those SDCMs by statistical tools and discovery methods available for simple SCMs. For example, one can apply adjustment criteria and Pearl's do-calculus (Forré and Mooij, 2019). Several causal discovery algorithms, originally designed for acyclic SCMs, like Local Causal Discovery (LCD) (Cooper, 1997), Y-structures (Mani, 2006), and the Fast Causal Inference (FCI) algorithm (Spirtes, Meek, and Richardson, 1999; Zhang, 2008; Mooij and Claassen, 2020), are directly applicable to simple SCMs as well (Mooij, Magliacane, and Claassen, 2020). Furthermore, the Joint Causal Inference (JCI) framework can be applied to combine data from different contexts (for example, observational and interventional) for causal discovery and inference purposes (Mooij, Magliacane, and Claassen, 2020).

Given that steady SDCMs for which all solutions equilibrate give rise to SCMs at equilibrium, the inverse problem becomes interesting as well: given an SCM, can we find an SDCM (with non-trivial dynamics) that equilibrates to this SCM and for which all solutions equilibrate? This question was answered affirmatively for a certain class of linear simple SCMs with additional constraints on the parameters by leveraging existing results from linear systems theory. We speculate that this result can be further generalized to allow for non-linearity. Perhaps surprisingly, this result allows to start from a second-order SDCM modeling a system of damped coupled harmonic oscillators, equilibrate it to obtain an SCM, and from that then construct a *first-order* SDCM with the same equilibrium SCM that describes all equilibrium states under any constant stochastic perfect intervention. This shows that the order of the dynamic structural equations is not necessarily constrained by the equilibrium SCM. Thus, the properties of the system at equilibrium may contain not enough information to identify the order of the dynamical equations.

We hope that the framework of SDCMs provides a natural starting point for modeling the causal mechanisms that underlie the dynamics of various systems, which could, in principle, be inferred from observations and experiments (see, for example, Bauer et al., 2017; Pfister, Bauer, and Peters, 2019; Liu et al., 2020). We believe that most of this work can also easily be adapted to discrete time by replacing the differential equations by difference equations. Future work might consist of (i) investigating the notion of local independence in SDCMs, (ii) studying how SDCM graphs can be interpreted causally, in particular if self-cycles or zeroth order equations are present, (iii) developing structure and parameter learning algorithms for SDCMs, and (iv) investigating possible extensions to stochastic dynamics by means of stochastic differential equations.

This appendix to Chapter 3 contains the proofs of all the theoretical results.

3.A PROOFS

Proof of Lemma 3.3.25. Let $\boldsymbol{X}_P : T \times \Omega \to \boldsymbol{\mathcal{X}}_P$ be a $C^{\boldsymbol{n}_P}$-stochastic process. For every $i \in I$ we can write the random differential equations

$$X_i^{(n_i)} = g_i(\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{X}_J, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P)$$

as a system of first-order random differential equations

$$\frac{d}{dt}\overline{X}_i^{(n_i-1)} = \widetilde{\boldsymbol{g}}_i(\overline{X}_i^{(n_i-1)}, \overline{\boldsymbol{X}}_{I\setminus i}^{(\boldsymbol{n}_{I\setminus i}-1)}, \boldsymbol{X}_J, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P)\,,$$

where $\widetilde{\boldsymbol{g}}_i : \mathcal{X}_i^{n_i} \times \boldsymbol{\mathcal{X}}_{I\setminus i}^{\boldsymbol{n}_{I\setminus i}} \times \boldsymbol{\mathcal{X}}_J \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \mathcal{X}_i^{n_i}$ is the mapping defined by

$$\widetilde{\boldsymbol{g}}_i(\bar{x}_i^{(n_i-1)}, \bar{\boldsymbol{x}}_{I\setminus i}^{(\boldsymbol{n}_{I\setminus i}-1)}, \boldsymbol{x}_J, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) := (x_i^{(1)}, \dots, x_i^{(n_i-1)}, g_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{x}_J, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P)).$$

Note that $\overline{X}_i^{(n_i-1)} = (X_i, X_i^{(1)}, \dots, X_i^{(n_i-1)})$ and $\frac{d}{dt}\overline{X}_i^{(n_i-1)} = (X_i^{(1)}, \dots, X_i^{(n_i-1)}, X_i^{(n_i)})$.

Substituting the functions $\boldsymbol{g}_J$ yields the following first-order RDE:

$$\frac{d}{dt}\overline{X}_i^{(n_i-1)} = \widetilde{\boldsymbol{g}}_i\big(\overline{X}_i^{(n_i-1)}, \overline{\boldsymbol{X}}_{I\setminus i}^{(\boldsymbol{n}_{I\setminus i}-1)}, \boldsymbol{g}_J(\overline{X}_i^{(n_i-1)}, \overline{\boldsymbol{X}}_{I\setminus i}^{(\boldsymbol{n}_{I\setminus i}-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P), \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P\big)\,.$$

Let $\widetilde{\boldsymbol{h}}_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) := \widetilde{\boldsymbol{g}}_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P)$. We can then write the dynamic SEs as:

$$\frac{d}{dt}\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)} = \widetilde{\boldsymbol{h}}_I\big(\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P\big)\,. \tag{3.9}$$

The assumed continuity of g_i and $\boldsymbol{g}_J$, the continuity of the exogenous process $\boldsymbol{E}_P$ and the assumption that $\boldsymbol{X}_P$ is a $C^{\boldsymbol{n}_P}$-stochastic process together imply that for almost all $\omega \in \Omega$ the function $(t, \bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}) \mapsto \widetilde{\boldsymbol{h}}_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}(t,\omega), \boldsymbol{E}_P(t,\omega))$ is continuous on $T \times \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I}$. Moreover, for each $\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)} \in \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I}$ the function $(\bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) \mapsto \widetilde{\boldsymbol{h}}_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P)$ is continuous and in particular measurable. Hence, for all $(t, \bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}) \in T \times \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I}$ the function $\omega \mapsto \widetilde{\boldsymbol{h}}_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}(t,\omega), \boldsymbol{E}_P(t,\omega))$ is $\mathcal{F}$-measurable.

Under the assumed condition, the following inequality holds for all $\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)} \in \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I}$, for all $\bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)} \in \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1}$ and for all $\boldsymbol{e}_P \in \boldsymbol{\mathcal{E}}_P$:

$$
\begin{aligned}
&\sum_{i\in I} \big\|\widetilde{\boldsymbol{h}}_i(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) - \widetilde{\boldsymbol{h}}_i(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P)\big\|^2 \\
&= \sum_{i\in I} \big\|\widetilde{\boldsymbol{g}}_i\big(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P\big) \\
&\qquad\qquad - \widetilde{\boldsymbol{g}}_i\big(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P\big)\big\|^2 \\
&= \sum_{i\in I} \Big[\|x_i^{(1)} - y_i^{(1)}\|^2 + \cdots + \|x_i^{(n_i-1)} - y_i^{(n_i-1)}\|^2 \\
&\qquad + \big\|g_i\big(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P\big) \\
&\qquad\qquad - g_i\big(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{g}_J(\bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P\big)\big\|^2\Big] \\
&\le \sum_{i\in I} \big[\|x_i^{(1)} - y_i^{(1)}\|^2 + \cdots + \|x_i^{(n_i-1)} - y_i^{(n_i-1)}\|^2 + \kappa^2\|x_i - y_i\|^2\big] \\
&\le (1+\kappa^2)\|\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)} - \bar{\boldsymbol{y}}_I^{(\boldsymbol{n}_I-1)}\|^2 .
\end{aligned}
$$

Hence the conditions of Theorem 1.2 in Bunke (1972) (or Theorem 3.2 in Neckel and Rupp (2013)) are satisfied, which proves that there exists an a.s. unique solution $\boldsymbol{X}_I$ of the system (3.9) of first-order RDEs for any partial initial condition $(t_0, \overline{\boldsymbol{X}}_{I,[0]}^{(\boldsymbol{n}_I-1)})$. Note that the solution $\boldsymbol{X}_I$ is a $C^{\boldsymbol{n}_I}$-stochastic process. Extend this to a solution $\boldsymbol{X}_{\mathcal{O}}$ on $\mathcal{O}$ by setting $\boldsymbol{X}_J = \boldsymbol{g}_J(\overline{\boldsymbol{X}}_I^{(\boldsymbol{n}_I-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P)$. The result satisfies the smoothness requirement; indeed, from the assumptions it follows for each $j \in J$ that X_j is a C^{n_j}-stochastic process. □

Proof of Proposition 3.3.26. Let $g_i : \mathcal{X}_i^{n_i} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}\setminus i} \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \mathcal{X}_i$ and $g_j : \boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \mathcal{X}_j$ for $i \in I$ and $j \in J$ be continuous mappings that make $\mathcal{R}$ satisfy Assumption 2-$(I \subseteq \mathcal{O})$. Consider the stochastic perfect intervention $\mathrm{do}(L, \boldsymbol{K}_L)$ with $L \subseteq \mathcal{O}$. Then, the mappings $h_i : \mathcal{X}_i^{n_i} \times \boldsymbol{\mathcal{X}}_{\mathcal{O}\setminus i} \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times (\boldsymbol{\mathcal{X}}_L \times \boldsymbol{\mathcal{E}}_P) \to \mathcal{X}_i$ for $i \in I \setminus L$ defined by

$$
h_i(\bar{x}_i^{(n_i-1)}, \boldsymbol{x}_{\mathcal{O}\setminus i}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, (\tilde{\boldsymbol{e}}_L, \boldsymbol{e}_P)) := g_i(\bar{x}_i^{(n_i-1)}, \boldsymbol{x}_{\mathcal{O}\setminus i}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P)
$$

and the mappings $h_j : \boldsymbol{\mathcal{X}}_{I\setminus L} \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times (\boldsymbol{\mathcal{X}}_L \times \boldsymbol{\mathcal{E}}_P) \to \mathcal{X}_j$ for $j \in \mathcal{O} \setminus (I \setminus L)$ defined by

$$
h_j(\boldsymbol{x}_{I\setminus L}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, (\tilde{\boldsymbol{e}}_L, \boldsymbol{e}_P)) := \begin{cases} g_j((\boldsymbol{x}_{I\setminus L}, \tilde{\boldsymbol{e}}_L), \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) & \text{if } j \notin L \\ \tilde{e}_j & \text{if } j \in L \end{cases}
$$

make $\mathcal{R}_{\mathrm{do}(L,\boldsymbol{K}_L)}$ satisfy Assumption 2-$(I \setminus L \subseteq \mathcal{O})$. □

Proof of Proposition 3.3.27. If the causal mechanism $f_{\mathcal{O}}$ is defined as in the proposition, then the mappings $g_I : \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I} \times \boldsymbol{\mathcal{X}}_J \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \boldsymbol{\mathcal{X}}_I$ and $g_J : \boldsymbol{\mathcal{X}}_I^{\boldsymbol{n}_I} \times \boldsymbol{\mathcal{X}}_P^{\boldsymbol{n}_P+1} \times \boldsymbol{\mathcal{E}}_P \to \boldsymbol{\mathcal{X}}_J$ are given by

$$\begin{aligned} g_I(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \boldsymbol{x}_J, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) = -B_{II^{(\boldsymbol{n}_I)}}^{-1}(B_{I\bar{I}^{(\boldsymbol{n}_I-1)}}\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)} - \boldsymbol{x}_I + B_{IJ}\boldsymbol{x}_J + B_{I\bar{P}^{(\boldsymbol{n}_P)}}\bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)} \\ + \Gamma_{IP}\boldsymbol{e}_P) \\ g_J(\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)}, \bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{e}_P) = -B_{JJ}^{-1}(B_{J\bar{I}^{(\boldsymbol{n}_I-1)}}\bar{\boldsymbol{x}}_I^{(\boldsymbol{n}_I-1)} + B_{J\bar{P}^{(\boldsymbol{n}_P)}}\bar{\boldsymbol{x}}_P^{(\boldsymbol{n}_P)} + \Gamma_{JP}\boldsymbol{e}_P)\,. \end{aligned}$$

The converse is shown by taking for $B_{II^{(\boldsymbol{n}_I)}}$ and B_{JJ} the identity matrices. □

Proof of Corollary 3.3.28. For a linear SDCM $\mathcal{R}$ that satisfies Assumption 1-($I \subseteq \mathcal{O}$) there always exists a $\kappa \in \mathbb{R}$ such that the uniformly-Lipschitz condition of Lemma 3.3.25 holds. □

Proof of Corollary 3.3.29. This follows directly from Corollary 3.3.28 and Proposition 3.3.26. □

Proof of Theorem 3.3.33. Let $\mathcal{G}_{[0]}^+ := \mathcal{G}_{[0]}^+(\mathcal{R})$ denote the augmented collapsed graph of $\mathcal{R}$. We can construct the a.s. unique global solution $\boldsymbol{X}$ of $\mathcal{R}$ by recursively substituting the solutions into each other along the topological ordering of the directed acyclic graph formed by the strongly connected components $S \subseteq \mathcal{I}$ of $\mathcal{G}_{[0]}^+$.

We construct an SCM that has $\mathcal{G}_{[0]}^+$ as its graph. Consider the SCM with endogenous variables X_i taking values in $\mathcal{C}^{n_i}(T, \mathcal{X}_i)$ for $i \in \mathcal{I}$, exogenous variables $\overline{X}_{[0],i}^{(n_i-1)}$ taking values in $\mathcal{X}_i^{n_i}$ for $i \in \mathcal{I}_{[0]}$, as well as exogenous variables E_j taking values in $\mathcal{C}^0(T, \mathcal{E}_j)$ for $j \in \mathcal{J}$. The structural equations of this SCM are taken to be of the following form. Let $S \subseteq \mathcal{I}$ be a strongly connected component of $\mathcal{G}_{[0]}^+$ and write $P := \mathrm{pa}_{\mathcal{G}_{[0]}^+}(S) \setminus S$. Observe that from Assumption 1-($I_S \subseteq S$) and $\mathcal{R}$ having a tight order tuple it follows that $\boldsymbol{n}_{J_S} = 0$ for $J_S = S \setminus I_S$. The structural equations for $j \in J_S$ are taken to be of the form:

$$X_j = g_j(\overline{\boldsymbol{X}}_{I_S}^{(\boldsymbol{n}_{I_S}-1)}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P). \tag{3.10}$$

For $i \in I_S$, we integrate the equation

$$X_i^{(n_i)} = g_i(\overline{\boldsymbol{X}}_{I_S}^{(\boldsymbol{n}_{I_S}-1)}, \boldsymbol{X}_{J_S}, \overline{\boldsymbol{X}}_P^{(\boldsymbol{n}_P)}, \boldsymbol{E}_P),$$

n_i times to turn it into

$$\begin{aligned}
X_i &= \iota(X^{(0)}_{[0],i}, X^{(1)}_i) \\
&= \iota(X^{(0)}_{[0],i}, \iota(X^{(1)}_{[0],i}, X^{(2)}_i)) \\
&= \iota(X^{(0)}_{[0],i}, \iota(X^{(1)}_{[0],i}, \iota(X^{(2)}_{[0],i}, X^{(3)}_i))) \\
&= \dots \\
&= \iota(X^{(0)}_{[0],i}, \iota(X^{(1)}_{[0],i}, \iota(X^{(2)}_{[0],i}, \quad \cdots \quad \iota(X^{(n_i-1)}_{[0],i}, X^{(n_i)}_i)))) \\
&= \iota(X^{(0)}_{[0],i}, \iota(X^{(1)}_{[0],i}, \iota(X^{(2)}_{[0],i}, \quad \cdots \quad \iota(X^{(n_i-1)}_{[0],i}, g_i(\overline{\boldsymbol{X}}^{(\boldsymbol{n}_{I_S}-1)}_{I_S}, \boldsymbol{X}_{J_S}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_P)}_P, \boldsymbol{E}_P))))) \\
&=: F_i(\overline{X}^{(n_i-1)}_{[0],i}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{I_S}-1)}_{I_S}, \boldsymbol{X}_{J_S}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_P)}_P, \boldsymbol{E}_P)
\end{aligned}$$

where we explicitly incorporate the initial conditions. The mapping $F_i : \mathcal{X}_i^{n_i} \times \mathcal{C}^{\boldsymbol{n}_{I_S \cup J_S \cup P}}(T, \boldsymbol{\mathcal{X}}_{I_S \cup J_S \cup P}) \times \mathcal{C}^0(T, \boldsymbol{\mathcal{E}}_P) \to \mathcal{C}^{n_i}(T, \mathcal{X}_i)$ defined in this way is continuous (being a composition of continuous mappings), and hence, measurable. The structural equations for $i \in I_S$ are then taken to be of the form

$$X_i = F_i(\overline{X}^{(n_i-1)}_{[0],i}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_{I_S}-1)}_{I_S}, \boldsymbol{X}_{J_S}, \overline{\boldsymbol{X}}^{(\boldsymbol{n}_P)}_P, \boldsymbol{E}_P). \tag{3.11}$$

The structural equations in (3.10) and (3.11) for all strongly connected components $S \subseteq \mathcal{I}$ of $\mathcal{G}^+_{[0]}$ together specify a well-defined SCM. Its exogenous variables $(\boldsymbol{E}_j)_{j \in \mathcal{J}}$ and $(\overline{X}^{(n_i-1)}_{[0],i})_{i \in \mathcal{I}_{[0]}}$ are assumed independent. The graph of the SCM is $\mathcal{G}^+_{[0]}$. As $\langle \mathcal{R}, I_S, S \rangle$ is assumed to be uniquely solvable, it follows that this SCM is uniquely solvable w.r.t. I_S. As this holds for every strongly connected component $S \subseteq \mathcal{I}$ of $\mathcal{G}^+_{[0]}$, the σ-separation Markov property (see Theorem 2.6.3.(2)) applies, proving the statement. □

Proof of Corollary 3.3.35. Extend the SCM constructed in the proof of Theorem 3.3.33 with endogenous variables $\overline{X}^{(n_i)}_{[1],i}$ taking values in $\mathcal{X}_i^{n_i+1}$ for $i \in \mathcal{I}_{[1]}$, and with the corresponding structural equations

$$\overline{X}^{(n_i)}_{[1],i} = \pi(X_i, \partial(X_i), \dots, \partial^{n_i}(X_i)) \,.$$

These functions are continuous, hence measurable, and therefore the extended SCM is well-defined with the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$ as its graph. The additional nodes are sink nodes that form their own strongly connected components, and the SCM is obviously also uniquely solvable w.r.t. each of these additional nodes. Hence, the σ-separation Markov property (see Theorem 2.6.3.(2)) holds for this extended SCM with graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$. □

Proof of Corollary 3.3.37. Observe that the transition graph $\mathcal{G}_{[0]\dots[1]}(\mathcal{R})$ is obtained from the evaluated augmented collapsed graph $\mathcal{G}^+_{[0]\dots[1]}(\mathcal{R})$ by graphically marginalizing out the nodes $\mathcal{I}$. The statement then follows from Lemma 3.3.2 in (Forré and Mooij, 2017), which states that σ-separations are preserved under graphical marginalization. □

Proof of Proposition 3.4.3. We show that if $\boldsymbol{X}$ is an equilibrating solution and $i \in \mathcal{I}$, then $\overline{X}_i^{(n_i)*} = (X_i^*, 0, \ldots, 0)$ almost surely. For all $0 \leq k_i \leq n_i$ we have for almost all $\omega \in \Omega$

$$\lim_{t\to\infty} X_i^{(k_i)}(t,\omega) = X_i^{(k_i)*}(\omega)\,.$$

Let $0 \leq m_i < n_i$. Let $\omega \in \Omega$ such that $\overline{X}_i^{(n_i)*}(t,\omega)$ converges. If $X_i^{(m_i+1)*}(\omega) > 0$, then there exists a $\bar{t} \in T$ such that $X_i^{(m_i+1)}(t,\omega) > \frac{1}{2}X_i^{(m_i+1)*}(\omega)$ for $t > \bar{t}$. From the mean value theorem, it follows that there exists a $c \in (\bar{t}, t)$ such that

$$X_i^{(m_i)}(t,\omega) - X_i^{(m_i)}(\bar{t},\omega) = X_i^{(m_i+1)}(c,\omega)(t-\bar{t}) > \tfrac{1}{2}X_i^{(m_i+1)*}(\omega)(t-\bar{t})$$

and hence $X_i^{(m_i)}(t,\omega)$ cannot converge to $X_i^{(m_i)*}(\omega)$. We get a similar contradiction under the assumption $X_i^{(m_i+1)*}(\omega) < 0$, and hence $X_i^{(m_i+1)*}(\omega) = 0$. We conclude that $\overline{X}_i^{(n_i)*} = (X_i^*, 0, \ldots, 0)$ almost surely. □

Proof of Proposition 3.4.4. We can rewrite the dynamic structural equations of $\mathcal{R}$ as

$$\begin{cases} \boldsymbol{X}_I' = -B_{II'}^{-1}(B_{II} - \mathbb{I}_I)\boldsymbol{X}_I - B_{II'}^{-1}B_{IJ}\boldsymbol{X}_J - B_{II'}^{-1}\Gamma_{I\mathcal{J}}\boldsymbol{E} \\ \boldsymbol{X}_J = -B_{JJ}^{-1}B_{JI}\boldsymbol{X}_I - B_{JJ}^{-1}\Gamma_{J\mathcal{J}}\boldsymbol{E}\,. \end{cases}$$

Eliminating $\boldsymbol{X}_J$ from the right-hand side by substitution yields the RDE

$$\boldsymbol{X}_I' = A\boldsymbol{X}_I + C\boldsymbol{E}\,,$$

where $A := B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}B_{JI} - B_{II} + \mathbb{I}_I)$ and $C := B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}\Gamma_{J\mathcal{J}} - \Gamma_{I\mathcal{J}})$. The matrix A is a Hurwitz matrix by assumption and thus invertible (note $\det(A) \neq 0$). The solutions of the ODE $\boldsymbol{x}' = A\boldsymbol{x} + C\boldsymbol{e}$, where the vector $\boldsymbol{e}$ does not depend on time, are of the form $\boldsymbol{x} = \exp(At)\boldsymbol{x}_0 - A^{-1}C\boldsymbol{e}$, where $\boldsymbol{x}_0$ is some vector. For any matrix A there exists a nonsingular matrix P (possibly complex) that transforms A into its Jordan normal form, that is, $P^{-1}AP = \Lambda$ is a block diagonal matrix where each block Λ_i is a Jordan block associated with the eigenvalue λ_i of A, and is a square matrix of order m_i of the form

$$\Lambda_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_i & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ \vdots & & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & \lambda_i \end{bmatrix}.$$

Therefore,

$$\begin{aligned}(\boldsymbol{X}_I)_t &= \exp(At)\boldsymbol{X}_{I,[0]} - A^{-1}C\boldsymbol{E}_t \\ &= \exp(P\Lambda P^{-1}t)\boldsymbol{X}_{I,[0]} - A^{-1}C\boldsymbol{E}_t \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} t^{j-1}\exp(\lambda_i t)R_{ij}\boldsymbol{X}_{I,[0]} - A^{-1}C\boldsymbol{E}_t\end{aligned}$$

with $\boldsymbol{X}_{I,[0]}$ some random variable, n the total number of block diagonal matrices, and the R_{ij}'s certain block matrices that depend on P and Λ (Khalil, 1996). Since A is a Hurwitz matrix by assumption and $\boldsymbol{E}$ is constant in time, we conclude that for all solutions $\boldsymbol{X}$ of $\mathcal{R}$,

$$\lim_{t\to\infty}(\boldsymbol{X}_I)_t = -A^{-1}C\boldsymbol{E}$$

and

$$\lim_{t\to\infty}(\boldsymbol{X}_J)_t = (B_{JJ}^{-1}B_{JI}A^{-1}C - B_{JJ}^{-1}\Gamma_{J\mathcal{J}})\boldsymbol{E}$$

almost surely.

At last, we consider replacing the condition that the exogenous process $\boldsymbol{E}$ is constant in time by the assumption that $\boldsymbol{E}$ may depend on time but is continuous, and that both $\boldsymbol{E}_t$ and $\exp(At)\int_{t_0}^{t}\exp(-As)C\boldsymbol{E}_s ds$ converge almost surely. Observe that the general solutions of $\boldsymbol{x}' = A\boldsymbol{x} + C\boldsymbol{e}$, where we allow $\boldsymbol{e}$ to be a time-dependent vector, are of the form $\boldsymbol{x} = \exp(At)\boldsymbol{x}_0 + \exp(At)\int_{t_0}^{t}\exp(-As)C\boldsymbol{E}_s ds$. Then, replacing the term $-A^{-1}C\boldsymbol{E}_t$ in the equation above for $(\boldsymbol{X}_I)_t$ by $\exp(At)\int_{t_0}^{t}\exp(-As)C\boldsymbol{E}_s ds$ implies also that $(\boldsymbol{X}_I)_t$ converges a.s., from which the result follows. □

Proof of Lemma 3.4.7. Let $\boldsymbol{X}$ be an equilibrating solution and let $\boldsymbol{E}$ converge a.s. to the random variable $\boldsymbol{E}^*$. Then

$$\boldsymbol{X}^* = \lim_{t\to\infty}\boldsymbol{X}_t = \lim_{t\to\infty}\boldsymbol{f}\big(\overline{\boldsymbol{X}}_t^{(\boldsymbol{n})}, \boldsymbol{E}_t\big) = \boldsymbol{f}\Big(\lim_{t\to\infty}\overline{\boldsymbol{X}}_t^{(\boldsymbol{n})}, \lim_{t\to\infty}\boldsymbol{E}_t\Big) = \boldsymbol{f}(\overline{\boldsymbol{X}}^{(\boldsymbol{n})*}, \boldsymbol{E}^*)$$

almost surely, where in the third equality we used the continuity of $\boldsymbol{f}$. □

Proof of Proposition 3.4.10. Consider the finite index set $\mathcal{J} = \{1, \dots, e\}$ for some $e \in \mathbb{N}$. The independence of $(\boldsymbol{E}_j)_{j\in\mathcal{J}}$ implies that, in particular, for every $t \in T$ the family of random variables $\tilde{\boldsymbol{E}} := \big((\boldsymbol{E}_j)_t\big)_{j\in\mathcal{J}}$ is independent, that is, we have $\mathbb{P}^{\tilde{\boldsymbol{E}}_t} = \prod_{j\in\mathcal{J}}\mathbb{P}^{(\boldsymbol{E}_j)_t}$, where $\tilde{\boldsymbol{E}}_t := \big((\boldsymbol{E}_1)_t, \dots, (\boldsymbol{E}_e)_t\big)$.

Because $\lim_{t\to\infty}\tilde{\boldsymbol{E}}_t = \lim_{\substack{n\to\infty \\ n\in\mathbb{N}}}\tilde{\boldsymbol{E}}_n$ a.s., we have $\lim_{\substack{n\to\infty \\ n\in\mathbb{N}}}\tilde{\boldsymbol{E}}_n = \tilde{\boldsymbol{E}}^*$ a.s., where $\tilde{\boldsymbol{E}}^* := (\boldsymbol{E}_1^*, \dots \boldsymbol{E}_e^*)$. This implies that $\tilde{\boldsymbol{E}}_n$ converges in distribution to $\tilde{\boldsymbol{E}}^*$ (see Remark 6.4 and Corollary 13.19 in Klenke (2014)), that is, the distribution of $\tilde{\boldsymbol{E}}_n$ converges weakly to the distribution of $\tilde{\boldsymbol{E}}^*$, that is, w-$\lim_{n\to\infty}\mathbb{P}^{\tilde{\boldsymbol{E}}_n} = \mathbb{P}^{\tilde{\boldsymbol{E}}^*}$.[35] Similarly, we have

35 Let $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \dots$ be probability distributions over $\mathbb{R}^d$, then $\mathbb{P}_n$ *converges weakly to* $\mathbb{P}$, denoted by w-$\lim_{n\to\infty}\mathbb{P}_n = \mathbb{P}$, if $\lim_{n\to\infty}\mathbb{P}_n(U) = \mathbb{P}(U)$ for all measurable sets U in $\mathbb{R}^d$ with $\mathbb{P}(\partial U) = 0$, where ∂U is the boundary of U, that is, the closure of U minus the interior of U.

$\text{w-lim}_{n\to\infty}\mathbb{P}^{(\boldsymbol{E}_j)_n} = \mathbb{P}^{\boldsymbol{E}_j^*}$ for every $j \in \mathcal{J}$. Applying Theorem 2.8 in Billingsley (1999) gives that

$$\mathbb{P}^{\tilde{\boldsymbol{E}}^*} = \text{w-lim}_{n\to\infty}\mathbb{P}^{\tilde{\boldsymbol{E}}_n} = \text{w-lim}_{n\to\infty}\prod_{j\in\mathcal{J}}\mathbb{P}^{(\boldsymbol{E}_j)_n} = \prod_{j\in\mathcal{J}}\mathbb{P}^{\boldsymbol{E}_j^*}\,.$$

We conclude that the family of random variables $(\boldsymbol{E}_j^*)_{j\in\mathcal{J}}$ is independent. □

Proof of Theorem 3.4.11. Let $\boldsymbol{X}$ be an equilibrating solution and let $\boldsymbol{E}$ converge a.s. to the random variable $\boldsymbol{E}^*$. From Lemma 3.4.7 it follows that

$$\boldsymbol{X}^* = \boldsymbol{f}(\overline{\boldsymbol{X}}^{(\boldsymbol{n})*}, \boldsymbol{E}^*) = \boldsymbol{f}(\bar{\boldsymbol{\imath}}(\boldsymbol{X}^*), \boldsymbol{E}^*) = \boldsymbol{f}^*(\boldsymbol{X}^*, \boldsymbol{E}^*) \quad \text{a.s.},$$

where we used in the second equality that $\bar{\boldsymbol{\imath}}(\boldsymbol{X}^*) = \overline{\boldsymbol{X}}^{(\boldsymbol{n})*}$, since for all $i \in \mathcal{I}$ we have that $\overline{X}_i^{(n_i)*}$ is a.s. equal to $(X_i^*, 0, \ldots, 0)$ by Proposition 3.4.3. □

Proof of Proposition 3.4.15. Suppose that the equilibrated SDCM $\mathcal{M}_{\mathcal{R}}$ has a solution $\boldsymbol{X}^*$. Then the stochastic process $\boldsymbol{X} : T \times \Omega \to \boldsymbol{\mathcal{X}}$ defined by $\boldsymbol{X}_t(\omega) := \boldsymbol{X}^*(\omega)$ is a solution of $\mathcal{R}$ that equilibrates to $\boldsymbol{X}^*$. □

Proof of Proposition 3.4.16. By definition, the graph of the equilibrated model $\mathcal{M}_{\mathcal{R}}$ has nodes $\mathcal{I} \subseteq \overline{\mathcal{I}}^{(\boldsymbol{n})}$ and the augmented graph of $\mathcal{M}_{\mathcal{R}}$ has nodes $\mathcal{I} \cup \mathcal{J} \subseteq \overline{\mathcal{I}}^{(\boldsymbol{n})} \cup \mathcal{J}$. For every $i \in \mathcal{I}$, a functional parent of i in $\mathcal{M}_{\mathcal{R}}$ is a functional parent in $\mathcal{R}$, since for all $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$ and for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ we have

$$x_i = f_i^*(\boldsymbol{x}, \boldsymbol{e}) \quad \Longrightarrow \quad x_i = f_i(\bar{\boldsymbol{\imath}}(\boldsymbol{x}), \boldsymbol{e})\,.$$

Note there are no integrated parents of i in $\mathcal{M}_{\mathcal{R}}$ and there are no functional parents of $j \in \mathcal{J}$. □

Proof of Theorem 3.4.18. This follows directly from Definitions 3.3.1, 3.3.7 and 3.4.9. One can easily check that

$$\begin{aligned}(\mathcal{M}_{\mathcal{R}})_{\text{do}(I,\boldsymbol{K}_I^*)} &= \langle \mathcal{I}, I \cup \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{E}}, \widetilde{\boldsymbol{f}^*}, (\boldsymbol{K}_I^*, \boldsymbol{E}^*)\rangle \\ &= \langle \mathcal{I}, I \cup \mathcal{J}, \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{E}}, \tilde{\boldsymbol{f}}^*, (\boldsymbol{K}_I, \boldsymbol{E})^*\rangle \\ &= \mathcal{M}_{\mathcal{R}_{\text{do}(I,\boldsymbol{K}_I)}}\,,\end{aligned}$$

where the intervened and equilibrated dynamic causal mechanism

$$\widetilde{\boldsymbol{f}^*} = \tilde{\boldsymbol{f}}^* : \boldsymbol{\mathcal{X}} \times (\boldsymbol{\mathcal{X}}_I \times \boldsymbol{\mathcal{E}}) \to \boldsymbol{\mathcal{X}}$$

is given by

$$\tilde{f}_i^*(\boldsymbol{x}, (\boldsymbol{e}_I, \boldsymbol{e}_{\mathcal{J}})) := \begin{cases} f_i(\bar{\boldsymbol{\imath}}(\boldsymbol{x}), \boldsymbol{e}_{\mathcal{J}}) & i \in \mathcal{I} \setminus I \\ e_i & i \in I\,. \end{cases}$$

□

Proof of Corollary 3.4.22. The statement follows immediately from Theorem 1 of Fisher and Fuller (1958) followed by application of Proposition 3.4.4.

Theorem 1 of Fisher and Fuller (1958) states that under the stated condition, there exists an invertible diagonal stabilization matrix $\Lambda \in \mathbb{R}^{d\times d}$ such that $-\Lambda^{-1}A$ is Hurwitz.[36]

Note first that by construction, $\mathcal{M}_{\mathcal{R}_{\mathcal{M},\Lambda}} = \mathcal{M}$. The SDCM $\mathcal{R}_{\mathcal{M},\Lambda}$ satisfies Assumption 1-($\mathcal{I} \subseteq \mathcal{I}$), that is, it can be written in the form of the equations in Proposition 3.4.4 with $I = \mathcal{I}$, where $B_{II'} = -\Lambda$ and $B_{II} = B$, and hence $B_{II'}^{-1}(-B_{II} + \mathbb{I}_I) = -\Lambda^{-1}A$ is Hurwitz. The statements now follow from Proposition 3.4.4. □

Proof of Corollary 3.4.23. The statement follows from Theorem 2.1 of Locatelli and Schiavoni (2012) followed by application of Proposition 3.4.4 and Theorems 3.4.11 and 3.4.18.

Theorem 2.1 of Locatelli and Schiavoni (2012) states that for every matrix $A \in \mathbb{R}^{d\times d}$ that satisfies for all subsets $I \subseteq \mathcal{I}$ the condition $\det(A_{II})\det(\operatorname{diag}(A_{II})) > 0$, there exists a diagonal matrix $D \in \mathbb{R}^{d\times d}$ such that the matrix $D_{II}A_{II}$ is Hurwitz for all $I \subseteq \mathcal{I}$. In particular, observe that this matrix D is invertible, since D_{II} is invertible for every $I \subseteq \mathcal{I}$ (note $\det(D_{II}) \neq 0$ due to $\det(D_{II}A_{II}) \neq 0$).

Let $\Lambda \in \mathbb{R}^{d\times d}$ be an invertible diagonal matrix such that $-\Lambda_{II}^{-1}A_{II}$ is Hurwitz for every $I \subseteq \mathcal{I}$. Note first that by construction, $\mathcal{M}_{\mathcal{R}_{\mathcal{M},\Lambda}} = \mathcal{M}$. Now let $\operatorname{do}(J, \boldsymbol{K}_J)$ be a stochastic perfect intervention for some subset $J \subseteq \mathcal{I}$ and $\boldsymbol{K}_J$ some stochastic process that is constant in time. The intervened SDCM $(\mathcal{R}_{\mathcal{M},\Lambda})_{\operatorname{do}(J,\boldsymbol{K}_J)}$ satisfies Assumption 1-($I \subseteq \mathcal{I}$) for $I := \mathcal{I} \setminus J$, that is, it can be written in the form of the equations in Proposition 3.4.4, where $B_{II'} = -\Lambda_{II}$, $B_{JJ} = -\mathbb{I}_{JJ}$, $B_{JI} = \mathbf{0}_{JI}$ the zero matrix and $\Gamma_{J\mathcal{J}}\boldsymbol{e} = \boldsymbol{K}_J$. Moreover,

$$B_{II'}^{-1}(B_{IJ}B_{JJ}^{-1}B_{JI} - B_{II} + \mathbb{I}_I) = -\Lambda_{II}^{-1}(\mathbb{I}_I - B_{II}) = -\Lambda_{II}^{-1}A_{II}\,,$$

which is Hurwitz, from which we conclude that every solution $\boldsymbol{X}$ of $(\mathcal{R}_{\mathcal{M},\Lambda})_{\operatorname{do}(J,\boldsymbol{K}_J)}$ is an equilibrating solution. Hence, from Theorem 3.4.11 it follows that for every solution $\boldsymbol{X}$ of $(\mathcal{R}_{\mathcal{M},\Lambda})_{\operatorname{do}(J,\boldsymbol{K}_J)}$, its limit $\boldsymbol{X}^*$ is a solution of the equilibrated model

$$\mathcal{M}_{((\mathcal{R}_{\mathcal{M},\Lambda})_{\operatorname{do}(J,\boldsymbol{K}_J)})} = (\mathcal{M}_{\mathcal{R}_{\mathcal{M},\Lambda}})_{\operatorname{do}(J,\boldsymbol{K}_J)} = \mathcal{M}_{\operatorname{do}(J,\boldsymbol{K}_J)}\,,$$

where we made use of Theorem 3.4.18. Note that $\boldsymbol{E}$ is assumed constant (in time), and hence $\mathcal{R}_{\mathcal{M},\Lambda}$ is steady; in addition, $\boldsymbol{K}_J$ is assumed to be constant. The solutions of $\mathcal{M}_{\operatorname{do}(J,\boldsymbol{K}_J)}$ are a.s. unique, because they satisfy the equations $\boldsymbol{X}_I^* = A_{II}^{-1}(B_{IJ}\boldsymbol{X}_J^* + \Gamma_{I\mathcal{J}}\boldsymbol{E})$ and $\boldsymbol{X}_J^* = \boldsymbol{K}_J$ almost surely. □

36 A simple counterexample of a system that cannot be stabilized in this way is given by taking the matrix

$$B = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

for which $\Lambda^{-1}(B - \mathbb{I})$ is not Hurwitz for any diagonal invertible matrix Λ.

4

CONCLUSION

The inception of structural causal models forms an important step in the understanding of causality. Their predictive power stems from the basic assumption that the causal mechanisms remain invariant when other mechanisms are subjected to intervention. This allows us to perform causal reasoning and draw conclusions on the effect of interventions, counterfactuals, and potential outcomes.

Learning such structural causal models often rests on the additional *assumption of acyclicity*. Although this assumption may seem like a reasonable approximation and can simplify the theoretical analysis, in practice, one is often interested in learning complex systems where feedback loops are present that cannot simply be ignored. Two common ways to deal with feedback are (1) assuming that the causal model is linear or has discrete-valued variables, and (2) "unrolling" the dynamics over time in terms of a dynamical system. By "unrolling" the dynamics over time, feedback loops are effectively removed. Typically, discrete methods based on learning dynamic Bayesian networks, structural equations models, or vector autoregressive models and nonlinear generalization of those are used in this setting.

Our main contribution is a general theory of statistical causal modeling for both structural causal models and dynamical systems suitable for modeling latent confounding, cyclic and nonlinear causal relationships. Our proposed solution establishes an important bridge between structural causal models and a large class of stochastic dynamical systems at equilibrium, allowing us to infer causal properties of stochastic dynamical systems by employing the statistical tools and discovery methods available for SCMs on equilibrium data.

To go beyond cyclic SCMs in the linear and discrete-valued variable cases, we proposed various notions of (unique) solvability that apply to SCMs without any of the former restrictions. These notions play a key role in extending many of the convenient properties of acyclic SCMs to the cyclic setting. We showed that these notions of (unique) solvability provided sufficient (and sometimes even necessary) conditions for the following convenient properties for SCMs: (i) it has a solution and/or induces a unique distribution over the variables; (ii) it has a marginalization on a subset of the variables; (iii) its marginalization respects the latent projection; (iv) it satisfies a Markov property and (v) its graph is consistent with the causal semantics.

We proposed simple SCMs, which are SCMs that are uniquely solvable w.r.t. every subset of the variables. This class extends the class of acyclic SCMs to the cyclic setting while having all the convenient properties (i)-(v). This answers research question 1 in the affirmative. One key property of simple SCMs is that their solutions always satisfy the conditional independencies implied by σ-separation. This allows one to directly extend many results and algorithms for acyclic SCMs to the more

general class of simple SCMs, by simply replacing d-separation with σ-separation. The class of simple SCMs forms a convenient and practical extension of the class of acyclic SCMs that is already used for causal modeling, reasoning, and learning.

To go beyond dynamical systems that are discretized over time, we proposed the framework of structural dynamical causal models (SDCMs) that enables modeling of stochasticity, time-dependence, and causality in a natural way. The SDCM framework can be seen as the stochastic-process version of the SCM framework that contains the classes of SCMs and random differential equations as special cases. We provided an equilibration operation for SDCMs that allows for the equilibration of an SDCM to an SCM such that the resulting SCM contains all the equilibrium solutions of the SDCM, without requiring any assumption on the number of equilibrium solutions. This answers research question 2 in the affirmative. The framework of SDCMs enables the modeling of arbitrary order differential equations, including zeroth-order equations. This allows us to model the equilibrium solutions of dynamical systems that were previously considered to fall outside their scope, such as the price, supply, and demand model of economics. Furthermore, the framework of SDCMs enables the modeling of stochasticity, which allows for modeling randomness in both the initial conditions and the parameters of the model.

The framework of SDCMs enables modeling of the causal semantics by associating a distinct causal dynamics mechanism to each observed process that can be changed independently of one another by stochastic interventions. We provided a graphical representation for SDCMs that is compatible with intervention and equilibration. This provides the basis for modeling the causal mechanisms that underlie the dynamics of the systems encountered in science and engineering. This answers research question 3 in the affirmative. We showed that the equilibration operation commutes with intervention which results in the preservation of the causal semantics under equilibration. Application of the equilibrium operation on SDCMs makes it possible to study the causal semantics of the equilibrium solutions by statistical tools and discovery methods available for SCMs. An interesting direction for further exploration is the comparison to the Markov ordering graph (Blom and Mooij, 2021), which encodes the conditional independencies of the equilibrated model that are not always described by the graph of the SCM (for example, in the case of perfect adaptation) which represents the functional relationships.

For the inverse problem of equilibration, we showed that one could construct a stable first-order SDCM that realizes the causal semantics of a linear SCM at equilibrium under certain conditions. This establishes a class of linear SCMs that model the causal equilibrium semantics of certain linear SDCMs. We showed that the properties of the system at equilibrium might not contain enough information to identify the order of the dynamical equations. For example, we showed that the equilibrium solutions of the damped harmonic oscillator, as a second-order SDCM, can be realized by a stable first-order linear SDCM. This result could potentially be further generalized to allow for non-linearities.

We proposed a Markov property for SDCMs with initial conditions, which is the first of its kind to the best of our knowledge. This Markov property is suitable for

both the solutions of the SDCM and the evaluation of the solutions at any point in time under certain conditions. This answers research question 4 in the affirmative. This Markov property holds under unique solvability conditions similar to those of SCMs. We provided sufficient conditions under which the existence and uniqueness of solutions for a given initial condition can be guaranteed.

Finally, we provided conditions for identifying directed paths and bidirected edges in the graph of an SCM. We showed that, in general, the presence or absence of a directed path and bidirected edge cannot always be identified. In the light of SDCMs, we provided an overview of possible causal explanations for the case that the directed paths and bidirected edges cannot be identified. This answers research question 5 in the affirmative. Furthermore, we showed that the counterintuitive behavior of "nonancestral" effects, that is, when an intervention on a variable may change the distribution of some of its nondescendants in the graph, in the equilibrated SCMs, can be explained by the dependence of the equilibrium states on different initial conditions. We showed that this could be viewed as selection bias due to equilibration.

BIBLIOGRAPHY

Aalen, Odd O. (1987). "Dynamic modelling and causality". In: *Scandinavian Actuarial Journal* 1987.3–4, pp. 177–190.

Ascher, Uri M. and Linda R. Petzold (1998). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. 1st. Philadelphia, USA: Society for Industrial and Applied Mathematics.

Balke, A. and J. Pearl (1994). "Probabilistic Evaluation of Counterfactual Queries". In: *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*. Vol. 1. Seattle, Washington, USA: AAAI Press, pp. 230–237.

Bauer, Stefan, Nico S Gorbach, Djordje Miladinovic, and Joachim M. Buhmann (2017). "Efficient and Flexible Inference for Stochastic Systems". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 6988–6998.

Beckers, Sander and Joseph Y. Halpern (2019). "Abstracting Causal Models". In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*. Vol. 33. 1. Honolulu, Hawaii, USA: AAAI Press, pp. 2678–2685.

Billingsley, Patrick (1999). *Convergence of Probability Measures*. 2nd. Wiley Series in Probability and Statistics: Probability and Statistics. New York, USA: John Wiley & Sons Inc.

Blom, Tineke, Stephan Bongers, and Joris M. Mooij (July 2019). "Beyond Structural Causal Models: Causal Constraints Models". In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI-19)*. Ed. by Ryan P. Adams and Vibhav Gogate. Tel Aviv, Israel: AUAI Press.

Blom, Tineke, Mirthe M. van Diepen, and Joris M. Mooij (2021). "Conditional Independences and Causal Relations implied by Sets of Equations". In: *Journal of Machine Learning Research* 22.178, pp. 1–62.

Blom, Tineke and Joris M. Mooij (2021). "Causality and independence in perfectly adapted dynamical systems". Preprint. Available at arXiv:2101.11885 [cs.AI].

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

Bongers, Stephan, Tineke Blom, and Joris M. Mooij (2022). "Causal Modeling of Dynamical Systems". Preprint. Available at arXiv:1803.08784v3 [cs.AI]. Submitted to the Journal of Causal Inference.

Bongers, Stephan, Patrick Forré, Jonas Peters, and Joris M. Mooij (2021). "Foundations of Structural Causal Models with Cycles and Latent Variables". In: *Annals of Statistics* 49.5, pp. 2885–2915.

Borovkov, A.A. (2013). *Probability Theory*. Universitext. Springer London.

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

Bunke, H. (1972). *Gewöhnliche Differentialgleichungen mit Zufälligen parametern*. Berlin, DE: Akademie-Verlag.

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

Coddington, Earl A. and Norman Levinson (1955). *Theory of Ordinary Differential Equations*. International series in pure and applied mathematics. New York, USA: McGraw-Hill.

Cohn, Donald L. (2013). *Measure Theory*. 2nd. Boston, USA: Birkhäuser.

Commenges, Daniel and Anne Gégout-Petit (2009). "A general dynamical statistical model with causal interpretation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.3, pp. 719–736.

Cooper, Gregory F. (1997). "A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships". In: *Data Mining and Knowledge Discovery* 1.2, pp. 203–224.

Dagum, Paul, Adam Galper, and Eric Horvitz (1992). "Dynamic Network Models for Forecasting". In: *Uncertainty in Artificial Intelligence*, pp. 41–48.

Dash, Denver (Jan. 2005). "Restructuring Dynamic Causal Systems in Equilibrium". In: *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. Ed. by Robert Cowell and Zoubin Ghahramani. Barbados: The Society for Artificial Intelligence and Statistics, pp. 81–88.

Dawid, A.P. (2002). "Influence Diagrams for Causal Modelling and Inference". In: *International Statistical Review* 70 (2), pp. 161–189.

Dehaene, S. (2020). *How We Learn: Why Brains Learn Better Than Any Machine ... for Now*. Penguin Publishing Group.

Didelez, Vanessa (2000). "Graphical models for event history analysis based on local independence". PhD thesis. Universität Dortmund.

Didelez, Vanessa (2007). "Graphical Models for Composable Finite Markov Processes". In: *Scandinavian Journal of Statistics* 34.1, pp. 169–185.

Didelez, Vanessa (Jan. 2008). "Graphical models for marked point processes based on local independence". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 245–264.

Didelez, Vanessa (2015). "Causal Reasoning for Events in Continuous Time: A Decision–Theoretic Approach". In: *Proceedings of the UAI 2015 Workshop on Advances in Causal Inference (UAI15-ACI)*. Ed. by Ricardo Silva, Ilya Shpitser, Robin J. Evans, Jonas Peters, and Tom Claassen. Vol. 1504. CEUR Workshop Proceedings. Amsterdam, The Netherlands: CEUR-WS.org, pp. 40–45.

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

Eaton, Daniel and Kevin Murphy (2007). "Exact Bayesian structure learning from uncertain interventions". In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. Ed. by Marina Meila and Xiaotong Shen. Vol. 2. Proceedings of Machine Learning Research. San Juan, Puerto Rico, pp. 107–114.

Eberhardt, F. (Aug. 2014). "Direct Causes and the Trouble with Soft Interventions". In: *Erkenntnis* 79.4, pp. 755–777.

Eberhardt, F., P. Hoyer, and R. Scheines (2010). "Combining Experiments to Discover Linear Cyclic Models with Latent Variables". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy, pp. 185–192.

Eberhardt, Frederick and Richard Scheines (2007). "Interventions and Causal Inference". In: *Philosophy of Science* 74.5, pp. 981–995.

Eichler, Michael (2007). "Granger causality and path diagrams for multivariate time series". In: *Journal of Econometrics* 137.2, pp. 334–353.

Eichler, Michael and Vanessa Didelez (July 2007). "Causal Reasoning in Graphical Time Series Models". In: *Proceedings of the Twenty-Third Conference on Uncertainty in*

Artificial Intelligence (UAI-07). Ed. by Ron Parr and Linda van der Gaag. Vancouver, BC, Canada: AUAI Press, pp. 109–116.

Evans, Robin J. (2016). "Graphs for Margins of Bayesian Networks". In: *Scandinavian Journal of Statistics* 43, pp. 625–648.

Evans, Robins J. (2018). "Margins of discrete Bayesian networks". In: *The Annals of Statistics* 46.6A, pp. 2623–2656.

Fisher, Franklin M. (1970). "A Correspondence Principle For Simultaneous Equation Models". In: *Econometrica* 38.1, pp. 73–92.

Fisher, Franklin M. (1972). "A Simple Proof of the Fisher-Fuller Theorem". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 71.3, pp. 523–525.

Fisher, Michael E. and A.T. Fuller (1958). "On the Stabilization of Matrices and the Convergence of Linear Iterative Processes". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54 (4), pp. 417–425.

Fisher, R.A. (1935). *The Design of Experiments*. The Design of Experiments. Oliver and Boyd.

Florens, Jean-Pierre and Denis Fougere (1996). "Noncausality in Continuous Time". In: *Econometrica* 64.5, pp. 1195–1212.

Forré, Patrick and Joris M. Mooij (Oct. 2017). "Markov Properties for Graphical Models with Cycles and Latent Variables". Preprint. Available at arXiv:1710.08775 [math.ST].

Forré, Patrick and Joris M. Mooij (Aug. 2018). "Constraint-based Causal Discovery for Non-Linear Structural Causal Models with Cycles and Latent Confounders". In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*. Ed. by Amir Globerson and Ricardo Silva. Monterey, CA, USA: AUAI Press.

Forré, Patrick and Joris M. Mooij (July 2019). "Causal Calculus in the Presence of Cycles, Latent Confounders and Selection Bias". In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI-19)*. Ed. by Ryan P. Adams and Vibhav Gogate. Tel Aviv, Israel: AUAI Press.

Foygel, Rina, Jan Draisma, and Mathias Drton (2012). "Half-trek Criterion for Generic Identifiability of Linear Structural Equation Models". In: *The Annals of Statistics* 40.3, pp. 1682–1713.

Friston, K.J., L. Harrison, and W. Penny (2003). "Dynamic causal modelling". In: *NeuroImage* 19.4, pp. 1273–1302.

Geiger, Dan (1990). *Graphoids: A Qualitative Framework for Probabilistic Inference*. Tech. rep. R-142. Los Angeles, USA: Computer Science Department, University of California.

Ghahramani, Zoubin (1998). "Learning Dynamic Bayesian Networks". In: *Adaptive Processing of Sequences and Data Structures*. Ed. by Gori M. Giles C.L. Vol. 1387. NN 1997. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 168–197.

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

Golub, G. and W. Kahan (1965). "Calculating the Singular Values and Pseudo-Inverse of a Matrix". In: *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2.2, pp. 205–224.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3, pp. 424–438.

Haavelmo, T. (Jan. 1943). "The Statistical Implications of a System of Simultaneous Equations". In: *Econometrica* 11.1, pp. 1–12.

Halpern, J. (July 1998). "Axiomatizing Causal Reasoning". In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*. Ed. by Gregory Cooper and Serafin Moral. Morgan Kaufmann, pp. 202–210.

Han, Xiaoying and Peter E. Kloeden (2017). *Random Ordinary Differential Equations and Their Numerical Solution*. Vol. 85. Probability Theory and Stochastic Modelling. Singapore: Springer.

Hansen, Niels and Alexander Sokol (2014). "Causal Interpretation of Stochastic Differential Equations". In: *Electronic Journal of Probabity* 19.

Hyttinen, A., F. Eberhardt, and P.O. Hoyer (Nov. 2012). "Learning Linear Cyclic Causal Models with Latent Variables". In: *Journal of Machine Learning Research* 13. Ed. by Christopher Meek, pp. 3387–3439.

Hyttinen, A., P.O. Hoyer, F. Eberhardt, and M. Järvisalo (Aug. 2013). "Discovering Cyclic Causal Models with Latent Variables: A General SAT-based Procedure". In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI-13)*. Ed. by Ann Nicholson and Padhraic Smyth. Bellevue, WA, USA: AUAI Press, pp. 301–310.

Iwasaki, Yumi and Herbert A. Simon (1994). "Causality and model abstraction". In: *Artificial Intelligence* 67.1, pp. 143–194.

Jentzen, Arnulf and Peter E. Kloeden (2011). *Taylor Approximations for Stochastic Partial Differential Equations*. Vol. 83. PCBMS-NSF regional conference series in applied mathematics. Philadelphia, USA: Society for Industrial and Applied Mathematics.

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

Khalil, Hassan K. (1996). *Nonlinear systems*. 2nd. Upper Saddle River, NJ, USA: Prentice Hall.

Klenke, Achim (2014). *Probability Theory. A Comprehensive Course*. 2nd. Universitext. London, GB: Springer.

Kloeden, P.E. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Vol. 23. Stochastic Modelling and Applied Probability. Berlin, DE: Springer.

Korb, Kevin B., Lucas R. Hope, Ann E. Nicholson, and Karl Axnick (2004). "Varieties of Causal Intervention". In: *PRICAI 2004: Trends in Artificial Intelligence*. Ed. by Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap. Berlin, DE: Springer, pp. 322–331.

Koster, J. T. A. (1999). "On the Validity of the Markov Interpretation of Path Diagrams of Gaussian Structural Equations Systems with Correlated Errors". In: *Scandinavian Journal of Statistics* 26.3, pp. 413–431.

Koster, J.T.A. (1996). "Markov Properties of Nonrecursive Causal Models". In: *The Annals of Statistics* 24.5, pp. 2148–2177.

Lacerda, Gustavo, Peter L. Spirtes, Joseph Ramsey, and Patrik O. Hoyer (July 2008). "Discovering cyclic causal models by independent components analysis". In: *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI-08)*. Ed. by David McAllester and Petri Myllymaki. AUAI Press, pp. 366–374.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman (2017). "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40, e253.

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

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

Lewis, David K. (1979). "Counterfactual Dependence and Time's Arrow". In: *Noûs* 13.4, pp. 455–476.

Liu, Junyu, Zichao Long, Ranran Wang, Jie Sun, and Bin Dong (2020). "RODE-Net: Learning Ordinary Differential Equations with Randomness from Data". Preprint. Available at arXiv:2006.02377 [math.NA].

Locatelli, Arturo and Nicola Schiavoni (2012). "A necessary and sufficient condition for the stabilisation of a matrix and its principal submatrices". In: *Linear Algebra and its Applications* 436.7, pp. 2311–2314.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin, Heidelberg: Springer.

Maathuis, M.H., D. Colombo, M. Kalisch, and P. Bühlmann (2009). "Estimating High-Dimensional Intervention Effects from Observational Data". In: *The Annals of Statistics* 37.6A, pp. 3133–3164.

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

Mason, S.J. (Sept. 1953). "Feedback Theory - Some Properties of Signal Flow Graphs". In: *Proceedings of the IRE*. Vol. 41. 9. IEEE, pp. 1144–1156.

Mason, S.J. (July 1956). "Feedback Theory - Further Properties of Signal Flow Graphs". In: *Proceedings of the IRE*. Vol. 44. 7. IEEE, pp. 920–926.

Meek, Christopher (Aug. 1995). "Strong Completeness and Faithfulness in Bayesian Networks". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*. Ed. by Philippe Besnard and Steve Hanks. Morgan Kaufmann, pp. 411–418.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill series in computer science. New York: McGraw-Hill.

Mogensen, Søren Wengel and Niels Richard Hansen (Feb. 2020). "Markov equivalence of marginalized local independence graphs". In: *Ann. Statist.* 48.1, pp. 539–559.

Mogensen, Søren Wengel, Daniel Malinsky, and Niels Richard Hansen (Aug. 2018). "Causal Learning for Partially Observed Stochastic Dynamical Systems". In: *Proceedings of the Thirty-Fourth conference on Uncertainty in Artificial Intelligence (UAI-18)*. Ed. by Amir Globerson and Ricardo Silva. Monterey, CA, USA: AUAI Press.

Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf (2016). "Distinguishing Cause from Effect using Observational Data: Methods and Benchmarks". In: *Journal of Machine Learning Research* 17.32. Ed. by Isabelle Guyon and Alexander Statnikov, pp. 1–102.

Mooij, Joris M. and Tom Claassen (Aug. 2020). "Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI-20)*. Ed. by Jonas Peters and David Sontag. Vol. 124. PMLR, pp. 1159–1168.

Mooij, Joris M. and Tom Heskes (Aug. 2013). "Cyclic Causal Discovery from Continuous Equilibrium Data". In: *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-13)*. Ed. by Ann Nicholson and Padhraic Smyth. AUAI Press, pp. 431–439.

Mooij, Joris M., Dominik Janzing, and Bernhard Schölkopf (Aug. 2013). "From Ordinary Differential Equations to Structural Causal Models: the deterministic case". In: *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI-13)*. Ed. by Ann Nicholson and Padhraic Smyth. Corvallis, Oregon, USA: AUAI Press, pp. 440–448.

Mooij, Joris M., Sara Magliacane, and Tom Claassen (2020). "Joint Causal Inference from Multiple Contexts". In: *Journal of Machine Learning Research* 21.99, pp. 1–108.

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

Neckel, Tobias and Florian Rupp (Dec. 2013). *Random Differential Equations in Scientific Computing*. Warsaw: Versita, De Gruyter publishing group.

Pearl, J. (July 1985). "A Constraint Propagation Approach to Probabilistic Reasoning". In: *Proceedings of the First Conference on Uncertainty in Artificial Intelligence (UAI-85)*. Ed. by Laveen Kanal and John Lemmer. AUAI Press, pp. 31–42.

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

Pearl, J. and R. Dechter (Aug. 1996). "Identifying Independence in Causal Graphs with Feedback". In: *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI-96)*. Ed. by Eric Horvitz and Finn Jensen. Morgan Kaufmann, pp. 420–426.

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

Penrose, R. (1955). "A generalized inverse for matrices". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 51.3, pp. 406–413.

Peters, J., D. Janzing, and B. Schölkopf (2013). "Causal Inference on Time Series using Restricted Structural Equation Models". In: *Advances in Neural Information Processing Systems 26*, pp. 154–162.

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

Peters, J., J. M. Mooij, D. Janzing, and B. Schölkopf (June 2014). "Causal Discovery with Continuous Additive Noise Models". In: *Journal of Machine Learning Research* 15. Ed. by Aapo Hyvärinen, pp. 2009–2053.

Peters, Jonas, Stefan Bauer, and Niklas Pfister (2020). "Causal models for dynamical systems". Preprint. Available at arXiv:2001.06208 [stat.ME].

Pfister, Niklas, Stefan Bauer, and Jonas Peters (2019). "Learning stable and predictive structures in kinetic systems". In: *Proceedings of the National Academy of Sciences* 116.51, pp. 25405–25411.

Rebane, George and Judea Pearl (1987). "The Recovery of Causal Poly-Trees from Statistical Data". In: *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence (UAI-87)*. Seattle, WA, pp. 222–228.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley, CA: University of California Press.

Richardson, T. (Mar. 2003). "Markov Properties for Acyclic Directed Mixed Graphs". In: *Scandinavian Journal of Statistics* 30.1, pp. 145–157.

Richardson, T.S. (Aug. 1996a). "A Discovery Algorithm for Directed Cyclic Graphs". In: *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI-96)*. Ed. by Eric Horvitz and Finn Jensen. Portland, OR, USA: Morgan Kaufmann, pp. 454–461.

Richardson, T.S. (Feb. 1996b). *Discovering Cyclic Causal Structure*. Tech. rep. CMU-PHIL-68. Carnegie Mellon University.

Richardson, T.S. and P. Spirtes (1999). "Automated Discovery of Linear Feedback Models". In: *Computation, Causation, and Discovery*. Ed. by C. Glymour and G. F. Cooper. Cambridge, MA: MIT Press, pp. 253–304.

Richardson, T.S. and P. Spirtes (2002). "Ancestral Graph Markov Models". In: *The Annals of Statistics* 30.4, pp. 962–1030.

Richardson, Thomas S. (1996c). "Models of Feedback: Interpretation and Discovery". PhD thesis. Carnegie Mellon University.

Richardson, Thomas S. and James Robins (2013). *Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality*. Tech. rep. 128. Center for Statistics and the Social Sciences.

Richardson, Thomas S. and James M. Robins (2014). "ACE Bounds; SEMs with Equilibrium Conditions". In: *Statistical Science* 29.3, pp. 363–366.

Roese, N. J. (1997). "Counterfactual Thinking". In: *Psychological Bulletin* 121 (1), pp. 133–148.

Rubenstein, Paul K., Stephan Bongers, Bernhard Schölkopf, and Joris M. Mooij (Aug. 2018). "From Deterministic ODEs to Dynamic Structural Causal Models". In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*. Ed. by Amir Globerson and Ricardo Silva. Monterey, CA, USA: AUAI Press.

Rubenstein, Paul K., Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf (Aug. 2017). "Causal Consistency of Structural Equation Models". In: *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI-17)*. Ed. by Gal Elidan and Kristian Kersting. Sydney, Australia: AUAI Press.

Rubin, Donald B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 66 (5), pp. 688–701.

Schölkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio (2021). "Toward Causal Representation Learning". In: *Proceedings of the IEEE* 109.5, pp. 612–634.

Schweder, Tore (1970). "Composable Markov Processes". In: *Journal of Applied Probability* 7.2, pp. 400–410.

Shpitser, Ilya and Judea Pearl (Sept. 2008). "Complete Identification Methods for the Causal Hierarchy". In: *Journal of Machine Learning Research* 9. Ed. by Peter Spirtes, pp. 1941–1979.

Sims, Christopher (1980). "Macroeconomics and Reality". In: *Econometrica* 48.1, pp. 1–48.

Sobczyk, K. (1991). *Stochastic Differential Equations. With Applications to Physics and Engineering*. 1st. Vol. 40. Mathematics and its Applications. Dordrecht, NL: Springer.

Soong, T.T. (1973). *Random Differential Equations in Science and Engineering*. Vol. 103. Mathematics in Science and Engineering. New York, USA: Academic Press.

Spelke, Elizabeth S. (1990). "Principles of Object Perception". In: *Cognitive Science* 14.1, pp. 29–56.

Spirtes, P. (Apr. 1993). *Directed Cyclic Graphs, Conditional Independence, and Non-recursive Linear Structural Equation Models*. Tech. rep. CMU-PHIL-35. Carnegie Mellon University.

Spirtes, P. (June 1994). *Conditional Independence in Directed Cyclic Graphical Models for Feedback*. Tech. rep. CMU-PHIL-54. Carnegie Mellon University.

Spirtes, P. (Aug. 1995). "Directed Cyclic Graphical Representations of Feedback Models". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*. Ed. by Philippe Besnard and Steve Hanks. Morgan Kaufmann, pp. 499–506.

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

Spirtes, P., T. Richardson, C. Meek, R. Scheines, and C. Glymour (1998). "Using Path Diagrams as a Structural Equation Modelling Tool". In: *Sociological Methods & Research* 27.2, pp. 182–225.

Spirtes, Peter, Christopher Meek, and Thomas S. Richardson (1999). "An Algorithm for Causal Inference in the Presence of Latent Variables and Selection Bias". In: *Computation, Causation and Discovery*. Ed. by Clark Glymour and Gregory F. Cooper. The MIT Press. Chap. 6, pp. 211–252.

Tian, Jin (Aug. 2002). *Studies in Causal Reasoning and Learning*. Tech. rep. R-309. Los Angeles, USA: Cognitive Systems Laboratory, University of California.

Tian, Jin and Judea Pearl (Aug. 2001). "Causal Discovery from Changes". In: *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI-01)*. Ed. by John Breese and Daphne Koller. Seattle, Washington, USA: Morgan Kaufmann, pp. 512–521.

Verma, T.S. (Jan. 1993). *Graphical Aspects of Causal Models*. Tech. rep. R-191. Los Angeles, USA: Computer Science Department, University of California.

Voortman, Mark, Denver Dash, and Marek J. Druzdzel (July 2010). "Learning Why Things Change: The Difference-Based Causality Learner". In: *Proceedings of the*

Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10). Ed. by Peter Grünwald and Peter Spirtes. Catalina Island, CA: AUAI Press, pp. 641–650.

White, Halbert (2006). "Time-series estimation of the effects of natural experiments". In: *Journal of Econometrics* 135.1, pp. 527–566.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies In Philosophy Of Science. New York: Oxford University Press.

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

Zhang, Jiji (2008). "On the Completeness of Orientation Rules for Causal Discovery in the Presence of Latent Confounders and Selection Bias". In: *Artificial Intelligence* 172.16-17, pp. 1873–1896.

Probability

Ω	Sample space	2
$\mathcal{F}$	σ-Algebra	2
$\mathcal{B}(\boldsymbol{\mathcal{X}})$	Borel σ-algebra of $\boldsymbol{\mathcal{X}}$	98
$\mathbb{P}$	Probability measure	2
$\mathbb{P}_{\mathcal{O}}$	Restricted probability measure on $\mathcal{O}$	31
$p(\boldsymbol{x})$	Probability density	2
$p(\boldsymbol{x} \mid \boldsymbol{y})$	Conditional probability density	3
$\perp\!\!\!\perp$	Independence symbol	3
a.s.	Almost surely	22

Graphs

$\mathcal{G}$	Directed (mixed) graph	51
$\mathcal{V}$	Set of nodes of a graph $\mathcal{G}$	51
$\mathcal{E}$	Set of directed edges of a graph $\mathcal{G}$	51
$\mathcal{B}$	Set of bidirected edges of a graph $\mathcal{G}$	51
$\mathcal{H}$	Set of hyperedges of a graph $\mathcal{G}$	61
π	Walk/path	52
$\mathrm{pa}_{\mathcal{G}}(\mathcal{U})$	Parents of nodes $\mathcal{U}$ in $\mathcal{G}$	52
$\mathrm{ch}_{\mathcal{G}}(\mathcal{U})$	Children of nodes $\mathcal{U}$ in $\mathcal{G}$	52
$\mathrm{an}_{\mathcal{G}}(\mathcal{U})$	Ancestors of nodes $\mathcal{U}$ in $\mathcal{G}$	52
$\mathrm{de}_{\mathcal{G}}(\mathcal{U})$	Descendants of nodes $\mathcal{U}$ in $\mathcal{G}$	52
$\mathrm{sc}_{\mathcal{G}}(\mathcal{U})$	Strongly connected components of nodes $\mathcal{U}$ in $\mathcal{G}$	52
$\mathcal{G}_{\mathcal{W}}$	Induced subgraph of $\mathcal{G}$ on nodes $\mathcal{W}$	51
$\mathcal{G}^{\mathrm{sc}}$	Graph of strongly connected components of $\mathcal{G}$	52
$\mathcal{G}^{\mathrm{acy}}$	Acyclified graph of $\mathcal{G}$	57
$\mathrm{do}(I)$	Perfect intervention on I	27
$\mathrm{twin}(\mathcal{I})$	Twin operation w.r.t. $\mathcal{I}$	28
$\mathrm{marg}(\mathcal{L})$	Latent projection/marginalization operation w.r.t. $\mathcal{L}$	41
acy	Acyclification operation	57
col	Collapsing operation	113

$\perp\!\!\!\perp_{\mathcal{G}}^{d}$	d-Separation in $\mathcal{G}$ symbol	54
$\perp\!\!\!\perp_{\mathcal{G}}^{\sigma}$	σ-Separation in $\mathcal{G}$ symbol	58

Structural causal models

$\mathcal{M}$	Structural causal model (SCM)	21
$\mathcal{I}$	Index set of endogenous variables	21
$\mathcal{J}$	Index set of exogenous variables	21
$\boldsymbol{\mathcal{X}}$	Space of endogenous variables	21
$\boldsymbol{\mathcal{E}}$	Space of exogenous variables	21
f	Causal mechanism	21
$\mathbb{P}_{\boldsymbol{\mathcal{E}}}$	Exogenous distribution	21
$\boldsymbol{E}$	Exogenous random variable	22
$\boldsymbol{X}$	Endogenous random variable	22
$\boldsymbol{X}_{\boldsymbol{\xi}_I}$	Potential outcome under the perfect intervention $\mathrm{do}(I, \boldsymbol{\xi}_I)$	49
$\mathbb{P}^{\boldsymbol{E}}$	Exogenous distribution associated to $\boldsymbol{E}$	22
$\mathbb{P}^{\boldsymbol{X}}$	Observational/interventional distribution associated to $\boldsymbol{X}$	28
$\mathbb{P}^{(\boldsymbol{X},\boldsymbol{X}')}$	Counterfactual distribution associated to $(\boldsymbol{X}, \boldsymbol{X}')$	29
$\equiv$	Equivalence relation	23
$\equiv_{\mathrm{obs}(\mathcal{O})}$	Observational equivalence relation w.r.t. $\mathcal{O}$	34
$\equiv_{\mathrm{int}(\mathcal{O})}$	Interventional equivalence relation w.r.t. $\mathcal{O}$	35
$\equiv_{\mathrm{cf}(\mathcal{O})}$	Counterfactual equivalence relation w.r.t. $\mathcal{O}$	35
$\mathcal{G}^a$	Augmented graph mapping	24
$\mathcal{G}$	Graph mapping	24
$\mathrm{do}(I, \boldsymbol{\xi}_I)$	Perfect intervention on I	26
twin	Twin operation	28
$\mathrm{marg}(\mathcal{L})$	Marginalization operation w.r.t. $\mathcal{L}$	39
acy	Acyclification operation	57
$\mathcal{M}_{\mathrm{do}(I,\boldsymbol{\xi}_I)}$	Intervened SCM	26
$\mathcal{M}^{\mathrm{twin}}$	Twin SCM	28
$\mathcal{M}_{\mathrm{marg}(\mathcal{L})}$	Marginal SCM	39
$\mathcal{M}^{\mathrm{acy}}$	Acyclified SCM	57
$\boldsymbol{g}_{\mathcal{O}}$	Measurable solution function w.r.t. $\mathcal{O}$	30

Stochastic processes

$\boldsymbol{X}'$	Derivative of $\boldsymbol{X}$	112
$\boldsymbol{X}^{(n)}$	n^{th}-Order derivative of $\boldsymbol{X}$	117

$\overline{\boldsymbol{X}}^{(\boldsymbol{n})}$	Complete $\boldsymbol{n}^{\text{th}}$-order derivative of $\boldsymbol{X}$	117
$C^{\boldsymbol{n}}$	Continuously $\boldsymbol{n}$-times differentiable	117

Structural dynamical causal models

T	Time interval	112
$\mathcal{R}$	Structural dynamical causal model (SDCM)	117
$\mathcal{I}$	Index set of endogenous processes	117
$\mathcal{J}$	Index set of exogenous processes	118
$\boldsymbol{\mathcal{X}}$	Space of endogenous processes	118
$\boldsymbol{\mathcal{E}}$	Space of exogenous processes	118
$\boldsymbol{n}$	Order vector	118
$\boldsymbol{f}$	Dynamic causal mechanism	118
$\boldsymbol{E}$	Exogenous stochastic process	118
$\boldsymbol{X}$	Endogenous stochastic process	118
$\overline{\mathcal{I}}^{(\boldsymbol{n})}$	Nodes of the complete $\boldsymbol{n}^{\text{th}}$-order derivative of $\boldsymbol{X}$	117
$\boldsymbol{X}^*$	Equilibrium state of $\boldsymbol{X}$	141
$\mathcal{M}_{\mathcal{R}}$	Equilibrated SDCM	143
$\mathcal{G}^a$	Augmented graph mapping	125
$\mathcal{G}$	Graph mapping	125
$\mathcal{G}^+_{[0]}$	Augmented collapsed graph mapping	138
$\mathcal{G}^+_{[0]\ldots[1]}$	Evaluated augmented collapsed graph mapping	139
$\mathcal{G}_{[0]\ldots[1]}$	Transition graph mapping	139
$\mathrm{do}(I, \boldsymbol{K}_I)$	Stochastic perfect intervention on I	121
$\mathcal{R}_{\mathrm{do}(I,\boldsymbol{K}_I)}$	Intervened SDCM	121
$(t_0, \overline{\boldsymbol{X}}^{(\boldsymbol{m}_I)}_{I,[0]})$	Initial condition	128

Quantifiers

$\exists \boldsymbol{x}$	"There exists an $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$" quantifier	102
$\forall \boldsymbol{x}$	"For all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$" quantifier	102
$\forall \boldsymbol{e}$	"For $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e} \in \boldsymbol{\mathcal{E}}$" quantifier	102

Others

$\emptyset$	Empty set	21
$\mathbf{1}$	Singleton	21
$\boldsymbol{n}$	Set $\{1, 2, \ldots, n\}$	22
$\mathbb{I}$	Identity matrix	68

A^+	Pseudoinverse of the matrix A	68
$pr_{\mathcal{E}}$	Projection mapping on $\mathcal{E}$	100

LIST OF PUBLICATIONS

This thesis is based on the following two publications:

- **Bongers**, **Stephan**, Patrick Forré, Jonas Peters, and Joris M. Mooij (2021). "Foundations of Structural Causal Models with Cycles and Latent Variables". In: *Annals of Statistics* 49.5, pp. 2885–2915.

- **Bongers**, **Stephan**, Tineke Blom, and Joris M. Mooij (2022). "Causal Modeling of Dynamical Systems". Preprint. Available at arXiv:1803.08784v3 [cs.AI]. Submitted to the Journal of Causal Inference.

The majority of ideas, text, figures, and experiments originated from the first author. All other authors had advisory roles and helped with discussing ideas and/or directly contributed in writing to a small number of sections. In particular, Joris Mooij had an important advisory role over all performed research.

The author has further contributed to the following publications:

- Rubenstein, Paul K., Sebastian Weichwald, **Stephan Bongers**, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf (Aug. 2017). "Causal Consistency of Structural Equation Models". In: *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI-17)*. Ed. by Gal Elidan and Kristian Kersting. Sydney, Australia: AUAI Press.

- Magliacane, Sara, Thijs van Ommen, Tom Claassen, **Stephan Bongers**, Philip Versteeg, and Joris M. Mooij (2018). "Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions". In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 10869–10879.

- Rubenstein, Paul K., **Stephan Bongers**, Bernhard Schölkopf, and Joris M. Mooij (Aug. 2018). "From Deterministic ODEs to Dynamic Structural Causal Models". In: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*. Ed. by Amir Globerson and Ricardo Silva. Monterey, CA, USA: AUAI Press.

- Blom, Tineke, **Stephan Bongers**, and Joris M. Mooij (July 2019). "Beyond Structural Causal Models: Causal Constraints Models". In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI-19)*. Ed. by Ryan P. Adams and Vibhav Gogate. Tel Aviv, Israel: AUAI Press.

SAMENVATTING – SUMMARY IN DUTCH

Mensen zijn opmerkelijk bekwaam om in een bepaalde situatie de relevante objecten en concepten te identificeren die ze gebuiken in het redeneren over welke acties er genomen dienen te worden. Men is hiertoe in staat zonder te beschikken over kennis van en inzicht in alle causale verbanden en natuurwetten die een rol spelen in de betreffende situatie. Een kind leert al op jonge leeftijd oorzaak en gevolg van elkaar te onderscheiden. Bijvoorbeeld wanneer het leert dat een glas stuk kan gaan wanneer je het een zetje geeft en het daardoor van de tafel valt. De kennis die voortvloeit uit een dergelijk incident krijgt in ons brein een conceptuele representatie waarmee men kan redeneren. Zo'n conceptuele representatie stelt ons in staat om snel nieuwe dingen te leren, welke we door de tijd heen steeds weten her te gebruiken en aan te passen aan nieuwe situaties. Het vermogen van de mens om problemen op te lossen door het hergebruiken en aanpassen van kennis en vaardigheden in nieuwe situaties vormt de basis van onze *intelligentie*.

In het veld *machine learning*, wat een tak is binnen *artificial intelligence* (AI), is een van de voornaamste doelen het ontwikkelen van machines die op bovengenoemde manier kennis en vaardigheden kunnen vergaren. Dit veld poogt met andere woorden systemen te bouwen die zelf nieuwe kennis en vaardigheden kunnen leren door ervaringen op te doen aan de hand van data. Vaak wordt dit gedaan door een (computer) model te leren uit geobserveerde data, in de hoop dat het geleerde model ook goede voorspellingen kan doen en beslissingen kan nemen bij het zien van nieuwe data. Ondanks het grote succes, werkt deze aanpak in de praktijk alleen als de kansverdeling van de nieuwe data niet te veel afwijkt van de kansverdeling van de 'getrainde' data. In het veld *causal learning* wordt ernaar gestreefd een causaal model (*causal model*) te leren dat goede voorspellingen kan doen en beslissingen kan nemen na bepaalde acties of interventies, waarbij de kansverdeling van de data naar aanleiding van de actie of interventie mogelijk af kan wijken van die van de getrainde data. In causal learning wordt gepoogd een modulaire representatie van de wereld te leren, waarin de modules de causale mechanismes representeren, die kunnen worden hergebruikt en aangepast in verschillende omstandigheden. Deze causale mechanismes beschrijven alle oorzaak-gevolg relaties in het causale model, die men kan weergeven in een causale graaf.

Een centraal probleem binnen causal learning is dat de gebruikte causale modellen vaak berusten op de aanname dat de causale graaf *acyclisch* is. Dat wil zeggen, dat er geen geen causale cycli (*feedback loops*) aanwezig zijn, oftewel, dat er geen causale *terugkoppeling* kan plaatsvinden. Alhoewel deze aanname misschien redelijk lijkt en vaak de theoretische analyse vereenvoudigt, kan causale terugkoppeling in de praktijk lang niet altijd worden genegeerd. Causale terugkoppeling is namelijk een veelvoorkomend fenomeen, bijvoorbeeld, in de natuur, in onze economie of in leersituaties. Door continue terugkoppeling met de omgeving houdt je lichaam

bijvoorbeeld een constante temperatuur; door terugkoppeling met de markt wordt de prijs van een product bepaald; en een student leert wat goed en fout is door de terugkoppeling die een docent geeft op een opdracht.

Het belangrijkste doel van dit proefschrift, getiteld *Causaal Modelleren & Dynamische Systemen: Een Nieuw Perspectief Op Terugkoppeling*, is het verder ontwikkelen van het veld *causal modeling* in de aanwezigheid van terugkoppeling. In hoofdstuk 1 geven we een introductie van de stof. In hoofdstuk 2 onderzoeken we hoe we terugkoppeling kunnen karakteriseren in *structural causal models* (SCM's), een klasse van causale modellen waar de causale mechanismes beschreven kunnen worden door middel van functionele relaties. In hoofdstuk 3 onderzoeken we hoe we terugkoppeling in stochastische dynamische systemen causaal kunnen interpreteren en hoe dit relateert aan SCM's in evenwicht. Tot op heden bestond er: 1) nog geen algemeen toepasbare theorie voor SCM's met cycli waar niet-lineaire relaties zijn toegestaan; en 2) nog geen brug tussen de wereld van stochastische dynamische systemen en SCM's in evenwicht, waarbij de stochastische dynamische systemen meerdere evenwichtstoestanden kunnen hebben. Deze brug is belangrijk, omdat dit het mogelijk maakt om eigenschappen van vele stochastische dynamische systemen in de wetenchap en engineering in evenwicht te bestuderen met de statistische technieken en leer methodes die beschikbaar zijn voor SCM's. In de volgende paragrafen lichten we onze bijdragen in meer detail toe.

In hoofdstuk 2 introduceren we een algemene theorie van statistisch causaal modelleren voor SCM's, waarin men latente confounding, cyclische en niet-lineaire causale relaties kan modelleren. We laten zien dat SCM's met cycli veel van de handige eigenschappen van *acyclische* SCM's niet hebben, zoals het bestaan van een (unieke) oplossing of van een zogeheten *Markov eigenschap*, welke nuttig is voor *causal learning*. We bewijzen dat onder bepaalde *solvabiliteits condities* veel van deze handige eigenschappen toch gelden voor SCM's in het algemeen. We introduceren een *marginalisatie operatie* voor SCM's, die gebruikt kan worden voor het verkrijgen van een marginaal model op een deel van de variabelen. Zo'n marginalisatie bestaat niet altijd voor SCM's met cycli zonder verdere aannames. We bewijzen dat deze marginalisatie operatie de waarschijnlijkheids en causale semantiek behoudt onder bepaalde lokale unieke solvabiliteits condities. Op een vergelijkbare wijze kunnen we de graaf van een SCM marginaliseren, wat we de *latente projectie* van de graaf noemen. We laten zien dat, in het algemeen, de marginalisatie van een SCM niet compatibel is met de latente projectie van zijn bijbehorende graaf, maar bewijzen dat dit wel zo is onder aanvullende lokale *ancestral* (voorouderlijke) unieke solvabiliteits condities.

Vervolgens introduceren we *simpele SCM's*. De klasse van simpele SCM's breidt de subklasse van acyclische SCM's uit naar het cyclische domein. Daarbij behouden ze veel handige eigenschappen, zoals het bestaan van unieke observationele en interventionele kansverdelingen, dat ze gesloten zijn onder interventie en marginalisatie, en dat ze een Markov eigenschap hebben. We illustreren dat de klasse van simpele SCM's een handige en praktische uitbreiding is van de klasse van acyclische SCM's die gebruikt kan worden voor *causal modeling*, *learning* en *reasoning*.

In hoofdstuk 3 introduceren we *structural dynamical causal models* (SDCM's). Deze klasse van modellen maakt het mogelijk om stochasticiteit, tijdsafhankelijkheid en causaliteit op een natuurlijke wijze samen te modelleren, en omvat de klassen van *structural causal models* (SCM's) en *random differential equations* (RDE's) als speciale gevallen. Een SDCM kan gezien worden als een stochastische proces versie van een SCM, waar de statische stochastische variabelen van het SCM worden vervangen door dynamische stochastische processen en hun afgeleiden. We geven een grafische representatie van SDCM's en geven condities voor het bestaan en de uniciteit van de oplossingen gegeven bepaalde beginvoorwaarden. Dit maakt het mogelijk om, onder bepaalde aannames, een *Markov eigenschap* voor SDCM's af te leiden, die gebruikt kan worden voor de oplossingen van het SDCM en voor de evaluatie van de oplossingen op elk moment in de tijd. We demonstreren dat SDCM's de basis vormen voor het modelleren van causale mechanismes die ten grondslag liggen aan de dynamica van systemen die men tegenkomt in de wetenschap en engineering.

Ten slotte introduceren we een *equilibratie operatie* voor SDCM's. Deze kan gebruikt worden voor het in evenwicht brengen van een SDCM naar een SCM, zodat de SCM alle de evenwichtstoestanden van het SDCM kan beschrijven, zonder enige aanname te doen over het aantal evenwichtstoestanden van het SDCM. Dit slaat een brug tussen SDCM's en SCM's in evenwicht en verschaft nieuw inzicht in de causale interpretaties van SCM's. Deze brug maakt het mogelijk om de causale semantiek te bestuderen van tal van stochastische dynamische systemen, inclusief degene met meerdere evenwichtstoestanden. Iets wat voorheen nog niet mogelijk was.

ACKNOWLEDGMENTS

I am grateful to all the people who have contributed in some way to the creation of this work.

First and foremost, I'd like to thank my advisor Prof. Joris Mooij, for giving me the opportunity to pursue a Ph.D. under his supervision and for introducing me to the fascinating field of causality. I appreciate the continuous support, the trust, and the encouragement he gave me to dive deep into the world of causality in all its encompassing mathematical intricacies. This work was not possible without the countless discussions and brainstorming sessions we had in front of the whiteboard. I could not have hoped for an advisor with greater clarity of thought, being a greater teacher, and being a greater guide to the unknown. I would like to thank Prof. Max Welling for co-advising me and for providing an excellent and supportive research environment.

I would like to thank my committee members, Judea Pearl, Thomas Richardson, Michel Mandjes, Frank Pijpers, Dominik Janzing, and Sara Magliacane, for the time and effort spent on reading my thesis. In particular, I would like to thank Judea Pearl, your work has been an indispensable source of inspiration to me and shaped my causal thinking from the start, and I would like to thank Thomas Richardson for all your excellent work from which I learned a great deal.

I am grateful to my colleagues in the causality group and the rest of the AMLab, the Amsterdam Machine Learning Lab. I would like to thank Sara Magliacane for making me feel at home in our causality group, Philip Versteeg for being my travel partner in crime to the many conferences and summer schools, and Thijs van Ommen for debugging my code. I also had the pleasure to collaborate with Tineke Blom, with whom I had tons of fruitful discussions and from whom I learned a great deal. Thank you, Noud de Kroon for all the climbing we did, where we could temporarily forget about all the scientific problems. In particular, I would like to thank Patrick Forré for being my partner in crime, who has guided me through the ups and downs of my Ph.D. I truly miss our endless conversations about math, science, academia, and the important things in life. I thank him for the support, encouragement, feedback, and advice he gave me throughout. He has been a role model for me, both within research and without, for which I am very grateful. Furthermore, I would like to thank Patrick Putzky, Peter O'Connor, Durk Kingma, Ted Meeds, Christos Louizos, and Tom Claassen for being great office mates. To all my other colleagues who enriched my life during my Ph.D. I thank Emiel Hoogeboom, Maurice Weiler, Shihan Wang, Zeynep Akata, Jakub Tomczak, Matthias Reisser and many more. Thank you Félice Arends, for all the administrative support.

I would also like to thank Bernhard Schölkopf, Jonas Peters, and Paul Rubenstein for hosting me at the Max Planck Institute in Tübingen during the spring of 2016.

It was a privilege and pleasure to brainstorm about causality and its relation to dynamical systems. I learned a great deal through my collaboration with Jonas Peters and Paul Rubenstein.

To my longtime friends Pouyesh and Sophie, Rick and Arti, Gijs and Femke, Renée, and Anne, thank you for all your support, for the much needed distractions, and for simply being in my life. To Christ, Ans and Floor, thank you for being such a warm and welcoming family. To Suus and Liset, my dear sisters, thank you for always catching me up when I fall during life or during climbing. To my parents, thank you for your unconditional love and support, for always being there for me, and for enduring my absence in the past few years. This thesis would not have existed without you.

Finally, I would like to thank Yke for walking by my side in life. Yke, your love, happiness, understanding, and support have been absolutely indispensable to me. I am eternally grateful to you for putting up with me and for starting our new chapter in life.