Path Dependent Structural Equation Models

Ranjani Srinivasan1, Jaron Lee2, Narges Ahmadi2,3, Ilya Shpitser2
1Electrical and Computer Engineering, Johns Hopkins University
2Computer Science, Johns Hopkins University
3Institute of Computational Biology, Helmholtz Munich Center
Corresponding author: Ranjani Srinivasan (ranjani@jhu.edu)

August 26, 2020

Abstract

Causal analyses of longitudinal data generally assume structure that is invariant over time. Graphical causal models describe these data using a single causal diagram repeated at every time step. In structured systems that transition between qualitatively different states in discrete time steps, such an approach is deficient on two fronts. First, time-varying variables may have state-specific causal relationships that need to be captured. Second, an intervention can result in state transitions downstream of the intervention different from those actually observed in the data. In other words, interventions may counterfactually alter the subsequent temporal evolution of the system. We introduce a generalization of causal graphical models, Path Dependent Structural Equation Models (PDSEMs), that can describe such systems. We show how causal inference may be performed in such models and illustrate its use in simulations and data obtained from a septoplasty surgical procedure.

1 Introduction

Many scientific questions and engineering tasks may only be approached by analyzing the behavior of a system over time. For example, in analysis of longitudinal studies in public health like the Nurses’ Health Study (NHS) [Belanger et al., 1978], repeated measurements of patients taken over time are used to understand long term impact of exposures and lifestyle choices on health and well-being. Similarly, tasks such as trajectory tracking, speech recognition and game playing require modeling the temporal evolution of a system.

Many models for longitudinal or time series data, such as hidden Markov models or Kalman filters, are graphical models, and most may be viewed as dynamic Bayesian networks (DBNs) [Murphy, 2012]. These models are used to predict the future evolution of systems, which is useful for game playing algorithms or object tracking, or find latent structures that best explain observations, which is useful in speech recognition problems, and phylogenetic analysis. Despite their complexity and usefulness, models of the above type deal with fundamentally associative relationships. However, understanding long term impact of exposures, lifestyle choices or policies, or providing decision support tools in complex domains that vary over time requires causal modeling. Models used in the literature for this task include graphical causal models [Pearl, 2000], marginal structural models [Robins, 1997], structural nested models [Robins, 1999], as well as models of counterfactual regret [Murphy, 2003] [Chakraborty and Moodie, 2013].

Causal models used for the analysis of longitudinal and time series data have generally assumed a repeated causal structure at each time point. For example, analysis of the impact of anti-retroviral therapy on the HIV infection progression in observational studies assumed the same variables relevant for the patient health and the same causal relationships linking them at each point in the study [Hernán et al., 2000]. Changes tracked over time are thus quantitative (such as the HIV developing resistance to the current drug), with the underlying causal structure remaining invariant over time. However, in addition to quantitative changes, many systems undergo qualitative changes as well, where observability, relevance, and causal relationships of variables change over time.

For example, a comprehensive study that follows the life course of people ought to take into account the fact that the person’s age, whether they are attending elementary school, high school, or university, and whether they are employed fundamentally changes the causal model describing many facts about that person. Similarly, economic and political models that track the evolution of unstable societies and “failed states” over time ought to take into account that society’s state during collapse of law and order, or during a civil war, or during a successful coup (where order is restored) fundamentally changes the causal model. Note that in all these examples states may recur: a failed state may suffer through a sequence of coups, while a person followed in a life course study may drop out of university, only to enroll again later. An additional feature of such system is that counterfactually different choices in the past may result in a qualitatively different evolution of the system from that point. For example, the life course of a person would likely be radically different had they, contrary to fact, not dropped out of college. Following the convention in the economics literature, we call this phenomenon path dependence [Liebowitz and Margolis, 2002]. Throughout this paper, we will use a relatively simple example of a system with path dependence: a surgery.
1.1 Septoplasty Surgery As a Structured System With Path Dependence

An average adult will undergo multiple surgical interventions in their life [Lee et al., 2008], many of them complex. In this paper, we will use septoplasty surgery as a running example. Septoplasty is a surgical procedure performed on the nasal cartilage, called the septum, to relieve nasal obstruction [Tajudeen and Kennedy, 2017]. See the Appendix for more details on this procedure.

For instructional and evaluation purposes, surgeries are often divided into discrete steps or "stages", each with its own intermediate goal [Ahmidi et al., 2015]. In septoplasty, these goals might include administering the anesthesia, making the initial incision, elevating the mucosal flap, reconstructing the cartilage, and so on [Fettman et al., 2009]. Each stage is associated with its own set of variables with relationships that may not be shared across stages. For example, stitching together a previously made incision is a routine task that may be executed by a surgical robot, while other tasks may require multiple tools, as well as skill and manual dexterity, and thus an experienced surgeon.

Another complication with surgeries is that procedures performed at a particular stage can and do go wrong, forcing surgeons to "double back" to previous stages of the surgery to correct mistakes, or deal with complications. Such a process may well occur multiple times. Surgeons may even be forced to perform additional procedures that were not within the scope of the original surgery. A surgery is generally performed by at least a pair of surgeons: a surgeon trainee, and an experienced attending surgeon. Which surgeon performs which stage can vary, with the need to train new surgeons being balanced with patient safety and operating costs. Perhaps unsurprisingly, the frequency with which previous surgery stages are revisited is related to which surgeon performs a particular stage of the surgery.

An approach appropriate for analyzing procedures such as septoplasty must be able to represent heterogeneous stages in a surgery, and the fact that stages may be revisited, perhaps multiple times. In addition, data obtained from prior surgeries may be used to assess counterfactual questions, such as what would the length of a surgery be had, contrary to fact, only experienced attending surgeons performed all procedures. Such questions can shed light on the causal impact of surgeon experience on outcomes such as quality of life, known to correlate with surgery length. Note that a counterfactual change of this sort may potentially alter the subsequent evolution of the system after the change, compared to the evolution actually observed in the data.

In this paper, we introduce the path-dependent structural equation model (PDSEM) for causal systems that exhibit qualitative changes over time, and path-dependence on counterfactual choices in the past, such as our septoplasty example. Our model can be viewed as a generalization of a causal dynamic Bayesian network that allows complex and repeating stage transitions, and distinct causal models at each stage, or as a generalization of a Markov decision process (MDP) where each state is modeled explicitly as a causal model, and where observed actions are not chosen by the optimizer as in reinforcement learning problems, but instead determined by the data generating process, possibly containing hidden variables.

As a result, our model combines complex and potentially looping state transitions of MDPs, and complex relationships among variables (such as confounding) of a causal model. PDSEMs may also be viewed as a generalization of a Markov chain endowed with graphical causal model semantics, which allows handling of confounding and analysis of counterfactual state transitions.

2 Background

2.1 Statistical and Causal Graphical Models

We introduce the necessary causal modeling ideas, before extending them to allow path-dependence.

A conditional graph $G(V, W)$ contains a set of random vertices $V$ and a set of fixed vertices $W$. Conditional graphs represent the structure of conditional distributions $p(V|W)$. We will consider conditional directed acyclic graphs (CDAGs) which contain directed edges, no directed cycles, and no edges with arrowheads into elements of $W$.

The statistical model corresponding to a CDAG is the set of distributions $p(V|W)$ such that $p(V|W) = \prod_{V \subseteq Y} p(V|pa_G(V))$, where $pa_G(V)$ are parents of $V$ in $G$. Such a distribution $p(V|W)$ is said to be Markov relative to $G(V, W)$. A CDAG where $W = \emptyset$ is called a directed acyclic graph (DAG), and yields the well-known Bayesian network model.

A causal model of a CDAG $G(V, W)$ is also a set of distributions, but on counterfactual random variables, representing causal relationships given a particular fixed context $W$. Given $Y \subseteq V$ and $A \subseteq V \setminus \{Y\}$, a counterfactual variable (given context $W$), written as $Y(a)|W$, represents the value of $Y$ in a hypothetical situation where $A$ were set to values $a$ by an intervention operation [Pearl, 2009].

In this work, we assume a natural generalization of Pearl’s functional model [Pearl, 2000] extended to a CDAG $G(V, W)$, where structural equations of the form $f_V(pa_G(V), \epsilon_V)$ determine $V(a_V)|W$. Here, $a_V \in \mathcal{X}_{\text{pa}_G(V)}$, with values in $\text{pa}_G(V) \cap W$ also serving as inputs to $f_V$. $\epsilon_V$ is a “noise” vari-
able that is exogenous and is the source of randomness in $V$.
We assume the joint distribution $p(\{v \mid v \in V\})$ factorizes as $\prod_{v \in V} p(v)$. That is, the counterfactual random variables $\{\{V_{a}^v \mid a_{V} \in X_{pa}(V) \} : V \in V\}$ are mutually independent [Pearl 2009].

A causal parameter of interest is said to be identified in a causal model if it is a function of the observed data distribution $p(V)$. Otherwise the parameter is said to be non-identified. In the functional model of a CDAG $G$ all intervention distributions are identified from $p(V|W)$.

**Lemma 1** For any $A \subseteq V$, $p(\{V \setminus A\}(a)|W)$ is identified from $p(V|W)$ in the functional model of $G(V|W)$ as

$$p(V \setminus A)(a)|W) = \prod_{v \in V \setminus A} p(V|pa_{G}(V))|_{A=a}. \tag{1}$$

(1) is a generalization of the g-formula [Robins 1986], and is a modified CDAG factorization with all terms corresponding to elements in $A$ missing, and elements in $A$ in remaining terms replaced by corresponding values in $a$.

### 2.2 Graphical Models In Discrete Time

While Bayesian networks lend themselves well to the modeling of static data, time series and dynamic data with temporal evolution proceeding in a series of discrete steps require more sophisticated models. Successful temporal models capture variable relationships not only within a particular point in time, but across time as well. A natural generalization of the Bayesian network model for discrete time temporal systems is the dynamic Bayesian network (DBN) model [Murphy 2012], defined as follows. Consider a set of vertices $V$, a DAG $G_1$ on $V$ (called a prior network), and a CDAG $G_{t+1}(V_{t+1}, V_t)$ on two time-indexed copies $V_t$ and $V_{t+1}$ (called a transition network). A DBN is defined by the pair of densities $p(V_1)$, $p(V_{t+1}|V_t)$ that factorize according to $G_1$ and $G_{t+1}$, respectively.

DBNs can represent a joint distribution representing the evolution of a multivariate system for any finite number of discrete time steps by “unrolling” the factorization of $p(V_1) \prod_{t=1}^{T-1} p(V_{t+1}|V_t)$ with respect to $G_1$ and $G_{t+1}$ as follows:

$$\left(\prod_{v \in V} p(V|pa_{G_1}(V))\right)^T \prod_{t=1}^{T-1} \left(\prod_{v \in V_{t+1}} p(V|pa_{G_{t+1}}(V))\right). \tag{2}$$

A DBN as defined here is first-order Markov, meaning that parents of any vertex can only occur in the current time point or the immediately prior time point.

The hidden Markov model (HMM) [Murphy 2012], is a popular special case of a DBN where the prior network contains two variables (a parent and a child) sometimes called the state variable and the sensor variable, and the transition network connects the state variables in the prior and the subsequent state. An example HMM is shown in Figure 1 with its (a) prior and (b) transition network as well as (c) time-unrolled graph for 4 steps.

The conditional probability distributions and associated parameters in HMMs (and DBNs more generally) are assumed to be time-invariant, which simplifies inference. The likelihood for a DBN may be obtained directly from its factorization. For instance, the likelihood for the HMM shown in Fig (b),(c) unrolled to 4 time steps is:

$$p(A_1; \eta_1) \prod_{t=1}^{4} p(L_t \mid A_t; \eta_t) \prod_{t=1}^{4} p(A_{t+1} \mid A_{t-1}; \eta_A).$$

Dynamic Bayesian networks can be generalized to causal models that can represent multivariate structured systems that evolve in time via a series of discrete steps. A natural approach to doing is to assume both prior and transition networks are causal DAGs and CDAGs, respectively, meaning that the value of every variable $V$ is determined in terms of its observed parents by means of a structural equation and a noise term $\epsilon_V$.

Thus, a causal DBN “unrolled” to a fixed set of time points $1, \ldots, T$ yields a standard causal DAG model with vertices $V_{1:T} \equiv V_1 \cup V_2 \cup \ldots \cup V_T$. In particular, for an intervention that sets $A \in V_{1:T}$ to constant values $a$, the interventional distribution $p(\{V_1:| A\}(a))$ is identified by a temporal generalization of the g-formula:

$$\prod_{v \in V \setminus A} p(V|pa_{G_1}(V)) \prod_{t=1}^{T-1} \prod_{v \in V_{t+1} \setminus A} p(V|pa_{G_{t+1}}(V)) \bigg|_{A=a} \tag{3}$$

Causal DBNs have been proposed in prior work. [Peters et al., 2013] illustrates how structural equations can be used in the context of time series data, addressing issues of identifiability. [Malinsky and Spirtes, 2018, 2019] present structure learning frameworks for causal dynamic networks and apply them to macroeconomic data.

A major modeling simplification employed by statistical and causal DBNs is that both structure and parameterization remain invariant over time, and thus temporal evolution proceeds along a set of states that are copies of each other. This makes DBNs ill-suited for capturing more complicated types of transition dynamics in a causal system. A popular model with simpler state structure but complex transition dynamics is known as a Markov Decision Process (MDP). A formal description of MDPs is given in the Appendix. However, MDPs are ill-suited for capturing complex causal relationships within a state, and even their generalizations such as partially observed MDPs (POMDPs) [Littman, 2009] are not able to handle completely unrestricted hidden variables. We now describe how path dependent structural equation models (PDESEMs) are able to capture both complex causal relationships within states, and complex transition dynamics across states.

### 3 Fully Observed PDESEMs

#### 3.1 A Simple PDESEM

To illustrate the structure of PDESEMs, we will use a simple example inspired by our surgery setting. We assume a surgery will consist of three states: $s^1$ (“incision”), the crucial state $s^2$ (“modification of bone/tissue”), and $s^3$ (“closing the incision”). We assume further that each state has the following relevant variables: $A$ (patient status prior to any procedures in
the current stage), B (whether the attending or resident surgeon is performing the procedure in the current stage), and C (the observed patient outcome for the stage after the procedure is performed).

The surgery always starts at $s^1$, and always concludes upon reaching the third state $s^3$. Procedures performed in the second state $s^2$ may either succeed, which will lead to $s^3$, or fail with some probability, leading the surgeon to revisit the first state $s^1$. The state transition diagram (of the sort employed for Markov chains) is shown in Fig. 2(b).

Relationships between variables in $s^1$ (viewed as the initial state) are shown by a causal diagram in Fig. 2(a). This graph corresponds to a standard set of structural equations as described in the previous section. This causal diagram, representing the state of the system at the start, serves the role played by the prior network in a causal DBN. Note that in addition to variables $A_1, B_1$ and $C_1$, this network contains the variable $S_1$, representing the state to transition to at time step 1. In general, the probability associated with this variable may depend on other variables in the current state, however in our simple model, the state $s^1$ transitions to $s^2$ with probability 1.

By analogy with causal DBNs, in addition to specifying the initial conditions of the model, we must also specify causal relationships involved in subsequent state transitions. This is accomplished by causal CDAGs for each possible state transition, shown in Fig. 2(c),(d),(e) respectively. We assume state spaces of variables associated with each state are the same across all state transition graphs and the prior graph, though variables themselves and their causal relationships may differ across graphs. For example, the state spaces of $A_1, B_1$ and $C_1$ in Fig. 2(a) and $A_{21}, B_{21}, C_{21}$ in Fig. 2(c) are the same, while the variables themselves (and the causal graphs relating them) are not. Random variables in each transition CDAG are indexed by both states in the transition, while fixed variables are indexed by the previous state in the transition. This is well-defined since variables in the previous state share state-spaces, regardless of the path taken to reach the previous state. It is this feature of PDSEMs that allows modeling of path dependence.

The model we describe represents a randomized controlled trial where the surgeon operating during the crucial second state $s^2$ is randomly assigned, hence $B_{12}$ in the transition graph into $s^2$ in Fig. 2(c) has no parents. Otherwise, we encode standard causal relationships we expect: C in the previous state influences $A, C$ in the next, and $A$ in the previous state influences $A$ in the next. Surgeon assignment $B_{12}$ in $s^2$ influences assignments in subsequent stages, whether they are $s^1$ or $s^3$. The state transition at $s^2$ depends on the outcome $C$ at that state. In $s^1, B$ does not influence $C$, since closing the incision is a relatively routine task that both more experienced and less experienced surgeons can adequately perform.

Like standard causal models, the PDSEM induces an observed data factorization and modified factorizations representing counterfactual situations. The observed data factorization of a PDSEM may be viewed as a combination of the factorization with the respect to a DAG of the starting state, and factorizations with respect to CDAGs representing state transitions, linked together by state transition probabilities (themselves functions of variables in those factorizations). As is the case with Markov chains, this factorization is not finite, but yields a well defined joint distribution $p_\infty$ over possible state trajectories, and associated state variable sets. In our case, the distribution $p_\infty$ factorizes as follows:

$$
\begin{align*}
    p_1 \prod_{t=1}^{\infty} (p_{12})^{1} (p_{23})^{1} (p_{21})^{1} (p_{3})^{1} (p_{3})^{1} \\
    p_1 = p(A_1) p(B_1 | A_1) p(C_1 | A_1, B_1) \tilde{p}(S_1) \\
    p_{12} = p(A_{12} | A_1, C_1) p(B_{12}) p(C_{12} | B_{12}, A_{12}, C_1) p(S_{12} | C_{12}) \\
    p_{23} = p(A_{23} | A_2, C_2) p(B_{23}) p(C_{23} | B_{23}, A_{23}) \tilde{p}(S_{23}) \\
    p_{21} = p(A_{21} | A_2, C_2) p(B_{21}) p(C_{21} | B_{21}, A_{21}) \tilde{p}(S_{21}),
\end{align*}
$$

where $s_t^i$ is the event “the state at time $t$ is $s^i$”, and all $\tilde{p}$ are deterministic by definition of our model.

The observed data distribution induced by the PDSEM may be viewed as a Markov chain represented by a set of graphical models. However, PDSEMs also allow us to reason about outcomes of interventions. For instance, assume we are interested in the situation where all procedures at every stage are performed, possibly contrary to fact, by the experienced attending surgeon (represented by value $b$). This entails consider the

![Figure 2: A simple PDSEM. (a) Causal structure of the initial state $S^1$. (b) The state transition diagram. (c),(d),(e) Causal diagrams representing possible transitions and subsequent states. (f) Causal relationships in a system evolving according to the state transitions: $s^1 \rightarrow s^2 \rightarrow s^3$. (g) A snapshot of a possible PDSEM trajectory represented as an unrolled DAG](image-url)
counterfactual joint distribution \( p_\infty(b) \) obtained from the counterfactual distribution \( p_1(b) = p(\{A_1, C_1, S_1\}) \) at the initial stage, and counterfactual distributions \( p_{12}(b), p_{23}(b), p_{21}(b) \) of the form \( p(\{A_{ij}, C_{ij}, S_{ij}\}) \{A_i, C_i, S_i\}) \) for each transition \( s^1 \rightarrow s^3 \). These counterfactual distributions are obtained by standard structural equation replacement semantics of interventions \cite{Pearl:2009}. The initial stage and transition probabilities are linked by state transition probabilities, as follows:

\[
p_{1}(b) \prod_{t=1}^{\infty} (p_{12}(b))^{i(s_{t-1}^{1}, s_{t}^{2})} (p_{23}(b))^{i(s_{t}^{2}, s_{t+1}^{3})} (p_{21}(b))^{i(s_{t+1}^{3}, s_{t+1}^{1})}
\]

Since the DAG representing the initial state, and all CDAGs representing transitions between states have standard structural equation semantics, the distribution \( p_\infty(b) \) is identified by using the g-formula for every component of the factorization of \( p_\infty \), yielding:

\[
p_0 \prod_{t=1}^{\infty} (p_{12})^{i(s_{t-1}^{1}, s_{t}^{2})} (p_{23})^{i(s_{t}^{2}, s_{t+1}^{3})} (p_{21})^{i(s_{t+1}^{3}, s_{t+1}^{1})}
\]

\[
p_1^* = p(A_1)p(C_1|A_1, b)p(S_1)
\]

\[
p_1^* = p(A_{12}|a_{12}, C_1)p(C_{12}|b, A_{12}, C_1)p(S_{12}|C_{12})
\]

\[
p_2^* = p(A_{23}|a_{23}, C_2)p(C_{23}|b, A_{23}, C_2)p(S_{23}|C_{23})
\]

\[
p_2^* = p(A_{21}|a_{21}, C_2)p(C_{21}|b, A_{21}, C_2)p(S_{21}|C_{21}).
\]

Note that while the distribution \( p(S_{12}|C_{12}) \) in the above factorization that governs how likely \( s^1 \) or \( s^3 \) are visited from \( s^2 \) remains the same, the probability that \( s^1 \) is visited from \( s^2 \) is lower in \( p_\infty(b) \) compared to \( p_\infty \). This is because \( B_{12} \) (a variable counterfactually set to \( b \)) causes \( C_{12} \), and \( C_{12} \) causes \( S_{12} \). This illustrates the ability of PDSEMs to encode counterfactually changing state transition probabilities from their observed values. This allows us to represent an intuitive feature of our example: surgeries where all stages are counterfactually performed by experienced attending surgeons will see many fewer returns to \( s^1 \) to correct mistakes.

### 3.2 Arbitrary PDSEM Models

An arbitrary PDSEM is defined using a set of states \( s \), with an initial state \( s^1 \) and an absorbing state \( s^3 \), a set \( T \) of state index pairs of the form \( (i, j) \), where \( s^1 \neq s^* \) representing allowed state transitions, a DAG \( G_1(V_1) \) for the initial state \( s^1 \), and for each \( (i, j) \in T \), a CDAG \( G_{ij}(V_{ij}, W_{ij}) \). Variables \( S_i \in V_1, \{S_{ij} : (i, j) \in T\} \) determine probabilities of transitioning from state to state. We assume \( S_i \), \( \{S_{ij} : (i, j) \in T\} \) have no outgoing edges. The DAG \( G_1 \), and CDAGs \( G_{ij} \) represent functional models for the initial state, and the appropriate state transitions, respectively. That is, in the initial state, each variable \( V \in V_1 \) is determined via \( f_V(p_{G_1}(V), e_V) \). Similarly, for each variable \( V \in V_{ij} \) in any state transition represented by \( G_{ij} \). These functional models have a restriction that for every state \( s^i \), any CDAG \( G_{ij} \) or DAG \( G_{j} \) will have random variables \( V_{ij}, V_{j} \) that share state spaces.

For conciseness, we will use the graphs \( G_1, \{G_{ij} : (i, j) \in T\} \) to represent the entire PDSEM.

Define \( V = V_1 \cup \bigcup_{(i, j) \in T} V_{ij} \). A PDSEM yields an observed data distribution \( p_\infty(V) \) with the following factorization:

\[
p_1(V_1) \prod_{t=1}^{\infty} \left( \prod_{(i, j) \in T} (p_{ij}(V_i|W_{ij}))^{i(s^t_{i-1}, s^t_j)} \right) ^{i(s^t_j)}
\]

\[
p_{ij}(V_{ij}|W_{ij}) = \prod_{V \in V_{ij}} p(V|pa_{ij}(V))
\]

An intervention in a PDSEM is defined on a set of treatment variables \( A_1 \subseteq V_1, A_{ij} \subseteq V_{ij} \) for each \( (i, j) \in T \), and corresponding values \( a = \{a_{1}\} \cup \{a_{ij} : (i, j) \in T\} \), a new counterfactual joint distribution \( p_\infty(V(a)) \), obtained from the counterfactual initial state distribution \( p_1(V_1(a_1)) \), and transition distributions \( p_{ij}(V_{ij}(a_{ij})|W_{ij} \) as:

\[
p_1(V_1(a_1)) \prod_{t=1}^{\infty} \left( \prod_{(i, j) \in T} (p_{ij}(V_i(a_{ij})|W_{ij}))^{i(s^t_{i-1}, s^t_j)} \right) ^{i(s^t_j)}
\]

Individual counterfactual distributions are obtained using standard structural equation replacement semantics.

Since the initial state and transitions are defined using structural equations, we obtain the following identification result, which generalizes the g-formula to PDSEMs.

**Lemma 2** Given a fully observed PDSEM, \( p_\infty(V(a)) \) is identified from \( p_\infty(V) \) as:

\[
p_1(V_1(a_1)) \prod_{V \in V_1\setminus A_1} p(V|pa_{ij}(V)) | A_{1}\to_{a}
\]

\[
p_{ij}(V_{ij}(a_{ij})|W_{ij}) \prod_{V \in V_{ij}\setminus A_{ij}} p(V|pa_{ij}(V)) | A_{ij}\to_{a}. \quad (4)
\]

### 4 Path Dependent Structural Equation Models With Hidden Variables

Before introducing hidden variable PDSEMs, we describe how hidden variables complicate standard causal models.

#### 4.1 Causal Inference With Hidden Variables

While interventional distributions are always identified (via the g-formula) in fully observed causal systems, identification theory becomes considerably more complicated in the presence of hidden variables. First interventional distributions may not be identified at all, and second identified distributions are equal to a modified version of a more complicated factorization than the DAG factorization. This nested factorization is associated with a special mixed graph derived from the hidden variable DAG.

To aid generalization of hidden variable causal models to hidden variable PDSEMs, we describe identification theory by
allowing a hidden variable causal model to depend on a set of fixed variables. Given a CDAG \( \mathcal{G}(V \cup H|W) \) where \( V \) correspond to observed random variables, and \( H \) to hidden random variables and \( W \) to fixed variables, define a conditional acyclic directed mixed graph (CADMG) \( \mathcal{G}(V|W) \) with directed (\( \rightarrow \)) and bidirected (\( \leftrightarrow \)) edges and vertices \( V, W \) as follows.

For every \( V_i, V_j \in V \cup W \), if there exists a directed path from \( V_i \) to \( V_j \) in \( \mathcal{G}(V \cup H|W) \) with all intermediate elements in \( H \), \( \mathcal{G}(V|W) \) contains an edge \( V_i \rightarrow V_j \). For every \( V_i, V_j \in V \), if there exists a path \( V_i \leftarrow \cdots \leftarrow V_j \) with no pair of adjacent edges on the path of the form \( \leftrightarrow \leftarrow \), with all intermediate elements in \( H \), \( \mathcal{G}(V|W) \) contains an edge \( V_i \leftrightarrow V_j \), \( \mathcal{G}(V|W) \) called the latent projection of \( \mathcal{G}(V \cup H|W) \). Identification theory implied by the causal model associated with a hidden variable DAG \( \mathcal{G}(V \cup H|W) \) may be phrased without loss of generality using \( \mathcal{G}(V|W) \).

If a conditional joint distribution \( p(V \cup H|W) \) is Markov relative to a CDAG \( \mathcal{G}(V \cup H|W) \), then the distribution \( p(V|W) \) obeys the nested factorization with respect to the latent projection CADMG \( \mathcal{G}(V|W) \). The Markov factorization of a CDAG \( \mathcal{G}(V|W) \) is phrased in terms of conditional distributions \( p(V|pa_G(V)) \), associated with every vertex \( V \in V \). The nested Markov factorization of a CADMG \( \mathcal{G}(V|W) \) is phrased in terms of Markov kernels \( q_\theta(S|pa^*_G(S)) \) (the set of strict parents \( pa^*_G(S) \) is defined as \( \bigcup_{i \in S} pa_G(S) \) \( \setminus S \)) associated with special subsets \( S \) of \( V \) called intrinsic sets. Each such intrinsic Markov kernel is a map from values of \( pa^*_G(S) \) to normalized densities over \( S \), and each is a functional of \( p(V) \) (which is not necessarily a conditional distribution).

The nested Markov factorization expresses \( p(V|W) \), and certain other distributions derived from \( p(V|W) \) as products of intrinsic Markov kernels. In particular, \( p(V|W) \) factorizes as \( \prod_{D} q_D(D) \left| pa^*_G(D) \right| \), where the product ranges over maximal bidirected connected sets \( D \) in \( \mathcal{G}(V|W) \) called districts [Tian and Pearl 2002]. The full details of the nested Markov factorization are in the Appendix.

It is known that any \( p(Y(a)|W) \) identified from \( p(V|W) \) given a causal model associated with a hidden variable CDAG \( \mathcal{G}(V \cup H|W) \) is equal to a functional of \( p(V|W) \) expressible as a modified nested Markov factorization. Specifically, we have:

\[
p(Y(a)|W) = \sum_{Y^\ast \setminus Y} \prod_{D \in D(Y^\ast)} q_D(D) \left| pa^*_G(D) \right| \big| A = a, \]

where \( Y^\ast \) is the set of ancestors of \( Y \) in \( \mathcal{G}(V|W) \) not through a variable in \( A \), the graph \( G_Y^\ast \) consists of vertices in \( Y^\ast \) and edges in \( \mathcal{G}(V|W) \) between \( Y^\ast \), and \( D(Y^\ast) \) is the set of districts in this graph. Each such district is always an intrinsic set in \( \mathcal{G}(V|W) \), and each corresponding \( q_D \) is a functional of \( p(V|W) \), yielding identification. Note that unlike many latent variable approaches [Rabiner 1989; Spearman 1950; Hohna et al. 2014], the theory presented makes no assumptions of any kind on \( H \), other than their location in the causal diagram.

### 4.2 PDSEMs With Latent Variables

The key observation in extending causal inference to latent variable PDSEMs is that as long as no transition CDAG depends on unobserved fixed context, the latent variable PDSEMs decompose into an initial state and a set of transitions such that causal inference results may be stated without loss of generality using latent projections of appropriate DAGs and CDAGs.

Fix a PDSEM defined given the initial state DAG \( \mathcal{G}(V_1 \cup H_1) \) and the set of transition CDAGS \( \mathcal{G}(V_{ij} \cup H_{ij}, W_i) \), for all \( (i, j) \in T \), such that \( S_1 \in V_1 \), \( S_{ij} \in V_{ij} \) for every \( (i, j) \in T \), for every \( j \) and all \( (i, j), (k, j) \in T \), \( H_{ij} = H_{kj} \) and \( V_{ij} = V_{kj} \). Moreover, for every \( j \), \( W_j \subseteq V_{ij} \) for any \( (i, j) \in T \). We assume the variables \( V \equiv \{V_1\} \cup \bigcup_{(i, j) \in T} V_{ij} \) are observed, while variables \( H \equiv \{H_1\} \cup \bigcup_{(i, j) \in T} H_{ij} \) are hidden.

Thus, the assumptions above may be rephrased as follows: state variables \( s^i, s^{i'} \) are always observed for any transition \( s^i \rightarrow s^{i'} \), every state has the same hidden and observed variables, regardless of current transition or initial state status, and all transitions depend only on observed variables in the previous state.

Given a latent variable PDSEM defined in this way, the observed data distribution \( p_\infty(V) \) is obtained from applying the usual transition probabilities to the margin at the initial state \( p(V_1) \equiv \sum_{H_1} p(V_1 \cup H_1) \), and the margins of all transition probabilities \( p(V_{ij}|W_i) \equiv \sum_{H_{ij}} p(V_{ij}|W_i \cup H_{ij}) \).

Define a set of treatments \( A \equiv \{A_1\} \cup \bigcup_{(i, j) \in T} A_{ij} \), where \( A_1 \subseteq V_1 \), \( A_{ij} \subseteq V_{ij} \) for all \( (i, j) \in T \), and the corresponding counterfactual distribution \( p_\infty(H(a) \cup V(a)) \) as in the previous section. Given the restrictions placed on the latent variable PDSEM, identification theory for \( p_\infty(V(a)) \) reduces to identification theory for \( p(V_1(1a_1)) \) in the latent projection ADMG \( \mathcal{G}_1(V_1) \), and \( p(V_{ij}(1a_{ij})|W_i) \) in the latent projection CADMG \( \mathcal{G}_{ij}(V_{ij}|W_i) \), as follows.

**Lemma 3** Given the latent variable PDSEM, \( p_\infty(V(a)) \) is identified from \( p_\infty(V) \) if and only if:

\[
p(\{V_1 \setminus A_1\}(a_1)) = \prod_{D \in D(V_1 \setminus A_1)} q_D(D) \left| pa^*_G(D) \right| ,
\]

with each kernel \( q_D(D) \left| pa^*_G(D) \right| \) above is identified from \( p_1(V_1) \), and for each \( (i, j) \in T \),

\[
p(\{V_{ij} \setminus A_{ij}\}(a_{ij})|W_i) = \prod_{D \in D(V_{ij} \setminus A_{ij})} q_D(D) \left| pa^*_G(D) \right| ,
\]

with each kernel \( q_D(D) \left| pa^*_G(D) \right| \) above is identified from \( p_{ij}(V_{ij}|W_i) \).

Identifying functionals for \( q_D \) are given in the Appendix.
5 Experiments

5.1 Simulations of PDSEM Without Latent Variables

We show how statistical inference may be performed in the example presented in Section 3.1 and Fig. 2. To recall, the system has three possible states \{s^1, s^2, s^3\} and three variables in each state \{A, B, C\}. For simplicity (and tractability in the latent variable model), we assume that the variables are all binary. Specifically, patient status prior to current stage is good (A = 1) or poor (A = 0), the surgeon is an experienced attending (B = 1) or a trainee (B = 0) and the outcome of current stage is good (C = 1) or poor (C = 0). State and transition DAGs are also identical to those in Fig. 2.

There are two sets of parameters associated with a generative model of this kind: \( p(S_{t+1} = s^j \mid S_t = s^i, V_t) \), where \( s^1_t \rightarrow s^2_{t+1} \) is a transition allowed by the model and \( p(V_{t+1} = v \mid S_{t+1} = s^j, S_t = s^i, V_t) \), where \( V^{ij} \in \{A^{ij}, B^{ij}, C^{ij}\}, v \in \{0, 1\} \) and, again, \( s^1_t \rightarrow s^2_{t+1} \) is an allowed transition. Parameters of this model are chosen to be reasonable for the surgery application, while yielding distribution Markov relative to appropriate graphs.

A dataset of \( N = 10000 \) “surgeries” (trials) were simulated, with all trials starting in the same state, with transitions generated according to the model, terminating at the absorbing state. Parameters of the PDSEM were estimated by maximum likelihood.

We used the PDSEM to consider the causal impact of experienced (attending) surgeons on average surgery length, versus a less experienced trainee surgeon. While imperfect, such an outcome is easy to measure, and is known to correlate with other measures of surgery quality, such as followup assessments of quality of life [Rambachan et al., 2013, Jackson et al., 2011]. We assessed this causal question by generating a set of sampled surgery trajectories where the attending performed in every state, and a set where the trainee performed in every state. These trajectories may be viewed as a Monte Carlo sampling scheme for evaluating the functional \( \Psi \). This comparison may be viewed as a generalization of the average causal effect (ACE) to PDSEMs.

The results are shown in Fig. 3. Surgeries performed by the attending are shorter \( (\mu = 3.36, \eta_{0.05} = 3, \eta_{0.95} = 5) \) than those performed by the trainee \( (\mu = 13.38, \eta_{0.05} = 3, \eta_{0.95} = 37) \) where \( \eta_p \) denotes the \( p^{th} \) percentile. Surgeries performed by the trainee have higher variance.

5.2 Data Application of the PDSEM

We now illustrate how PDSEM may be applied to analyze data obtained from a surgery. The dataset we chose consists of 236 septoplasty procedures conducted at our institution’s research hospital. 57343 timestamped records were collected regarding tool and personnel activity. Surgeries consist of six distinct phases - \( s^1 \) (opening of the septum), \( s^2 \) (raising septal flaps), \( s^3 \) (removal of deviated septal cartilage and bone), \( s^4 \) (reconstruction), \( s^5 \) (closing of the incision), and \( s^6 \) (other activity). In addition, there is an artificial absorbing state \( s^\text{end} \), representing the end of the procedure. Procedures are often led by an attending, with a surgeon trainee assisting. Of the surgeries, 42.79% of them were performed fully by the leading attending; the others were performed by a team. Additionally, attending surgeons perform the procedure for 64.98% of all operating time and trainees the rest. Twelve different surgical tools were tracked for use. Each phase of the surgery requires different techniques and tools. The progression of the surgery through the distinct phases is not monotonic - it is common for surgeons to return to phases already visited. The state transition diagram representing allowed state transitions is presented in Fig. 5. We chose to discretize all variables into two categories. Model parameters were estimated by maximum likelihood. More details about the data and model can be found in the Appendix.

As before, we considered the causal impact of surgeon experience on average length of surgery, evaluated by consider counterfactual trajectories and comparing to those actually observed in the data. Estimation of \( p(s_t \mid s_{t-1}, v_{t-1}) \) at all levels of \( s_{t-1}, v_{t-1} \) is not always possible due to finite sample limitations. To address this, we apply additive smoothing to \( p(s_t \mid s_{t-1}, v_{t-1}) \), based on the empirical distribution \( p(s_t \mid s_{t-1}) \). Goodness of fit is illustrated in Fig. 6 and results are presented in Fig. 7. We have made considerable assumptions in modeling our PDSEM and have closely matched the generative model to the empirical distribution (Fig. 4). We observe that the causal effect of surgeon skill on surgery length, given parameters learned in our data, is close to zero. This indicates that policies that govern the trade-off between the need to...
Figure 5: Histograms of hypothetical surgeries performed only by a junior trainee surgeon (blue) versus hypothetical surgeries performed only by a senior attending surgeon (orange). Surgeries performed by the attending are slightly longer (µ = 244.3.91, σ = 139.9) than those of the trainee (µ = 233.5, σ = 125.9).

Figure 6: The state transition diagram for the surgery data application.

train surgeons, and overall surgery quality (as quantified by our chosen outcome) are effective at our institution.

In our analysis, we have assumed a fully observed PDSEM, meaning that statistical inference may be directly adapted from Bayesian network models. Generalizing statistical inference in PDSEMs with hidden variables to likelihoods based on parameterizations of the nested Markov model [Richardson et al., 2017] presents a number of open problems, which we discuss in the Appendix.

6 Conclusions

In this paper, we have introduced the Path Dependent Structural Equation Model (PDSEM) for longitudinal data which unifies complex state structure from DBNs and complex state transition dynamics from MDPs. It can also be seen as a graphical model generalizing the dynamics of a Markov chain with state-specific dynamics. We have described counterfactuals associated with these causal models that can alter the subsequent temporal evolution of the system, identification theory for such counterfactuals in terms of the observed data distribution, and described estimation. We showed the utility of the model in clinical settings using simulations as well as real data from a septoplasty surgical procedure. Developing novel methods for efficient Monte Carlo sampling based statistical inference for hidden variable versions of PDSEMs based on the nested Markov model is a promising area of future work.
A The Septoplasty Surgical Procedure, and its PDSEM Model

Septoplasty is a surgical procedure performed on the nasal cartilage, called the septum, to relieve nasal obstruction [Tajudeen and Kennedy 2017]. A deviated or deformed septum is the most common cause of such an obstruction. Apart from nasal obstruction, a significantly deviated nasal septum has also been implicated in epistaxis, sinusitis, obstructive sleep apnea, and headaches which can act as diagnosis factors. The procedure involves cartilage resection, modification or a graft. The outcome of septoplasty is typically a score/index constructed from a questionnaire investigating quality of life measures and perceived nasal obstruction levels, like Nasal Obstruction Septoplasty Effectiveness (NOSE) and the Fairley Nasal Questionnaire (FNQ) [Fettman et al. 2009].

For instructional and evaluation purposes, surgeries are of- ten divided into discrete steps or "stages", each with its own intermediate goal [Ahmadi et al. 2015]. Our data from the septoplasty procedure was manually annotated by clinical experts and divided into the following stages:

- $s_1$: opening of the septum,
- $s_2$: raising septal flaps,
- $s_3$: removal of deviated septal cartilage and bone,
- $s_4$: reconstruction,
- $s_5$: closing of the incision,
- $s_6$: activity not otherwise included in the above 5 phases,
- $s_{\text{end}}$: end of surgery state (which contains no variables).

The variables in our data are the following:

$V = \{K: \text{knife}, C_1: \text{cottle}, D_1: \text{short needle driver}, D_2: \text{long needle driver}, G: \text{gorney scissors}, O: \text{other tools}, C_2: \text{suction cannula}, M: \text{main surgeon exists}, S: \text{suction exists}, A_1: \text{main surgeon is an attending}, A_2: \text{suction done by attending}, T: \text{duration of that phase is greater than 10 seconds}\}$

- $V_{s_1} = \{K, O, C_2, M, S, A_1, A_2, T\}$,
- $V_{s_2} = \{K, C_1, O, C_2, M, S, A_1, A_2, T\}$,
- $V_{s_3} = \{K, C_1, D_1, D_2, O, C_2, M, S, A_1, A_2, T\}$,
- $V_{s_4} = \{K, C_1, G, O, C_2, M, S, A_1, A_2, T\}$,
- $V_{s_5} = \{D_1, D_2, O, C_2, M, S, A_1, A_2, T\}$,
- $V_{s_6} = \{K, C_1, O, C_2, M, S, A_1, A_2, T\}$.

State DAGs were determined based on clinician recommendation. Parents of each variable in any state are the exact same variable in the previous state (and time point), if it exists, aside from the parents in the same time point indicated by state DAGs. For some $v_i \in V_j$, $\text{pa}_G(v_i) = \{v_i : \text{name}(v_i) = \text{name}(v_j), v_i \in V_i\}$, where $\text{name}(v_j) = v$.

Allowed state transitions were determined based on observed state transitions $V_i \rightarrow V_j$, subject to some thresholding criteria (needed to have had at least 5 observed transitions in data). The permitted state transitions are summarized in Figure 6 in the main paper.

B Markov Decision Processes

In a finite MDP, an agent and environment interact at discrete time steps $t = 0, 1, \ldots, T$, with the agent observing the environment in state $V_t$, taking action $A_t$, to land in state $V_{t+1}$, receiving a reward $R_{t+1}$ [Sutton and Barto 2018]. A finite MDP is defined by the tuple $(V, A, R, p(V_{t+1} = v', R_{t+1} = r|V_t = v, A_t = a), \gamma)$. Where $V$ is a finite set of states, $A$ is a set of actions, $R$ is a set of rewards, $p(V_{t+1} = v', R_{t+1} = r|V_t = v, A_t = a)$ is the probability of moving from state $V_t = v$ while taking action $A_t = a$ to the state $V_{t+1} = v'$, and getting reward $R_{t+1} = r$, and $0 \leq \gamma \leq 1$ is a discount factor that represents diminishing importance of future rewards. A policy $\pi(a|v) : V \rightarrow A$ is a map that represents the probability of taking an action $a$ in state $v$. Policies are often deterministic, mapping each state to a specific action. Under a policy $\pi$, we define value $G_{\pi}(v)$ of a state $v$ as the expected cumulative reward, starting at state $v$ and following $\pi(a|v)$ thereafter. $G_{\pi}(v)$ can be written in the form of a recursive equation as follows:

$$G_{\pi}(v) = E_v\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | V_t = v \right]$$

$$= \sum_a \pi(a|v) \sum_{s'} \sum_r p(s', r|s, a) \left[ r + \gamma G_{\pi}(v') \right]$$ (6)

This can be viewed as a consistency condition between value $G_{\pi}(v)$ and value $G_{\pi}(v')$ of possible successor state. Since value functions define a partial ordering over policies, we have $\pi \geq \pi'$ if and only if $G_{\pi}(v) \geq G_{\pi'}(v)$ for all $v \in V$. The optimal policy $\pi^*$ may not be unique, and has the optimal value function: $G^*(v) = \max_{\pi} G_{\pi}(v)$ for all $v \in V$.

An important special case: Consider an MDP with the following features: First, there are absorbing states $V^* = \{v_1^*, \ldots, v_b^*\}$. Second, the reward is non-zero only if there is a transition from a non-absorbing state $v$ to absorbing state $v^*$. That is, $R_t = 0$ if $v_t \not\in V^*$ or if $v_{t-1} \in V^*$, and $R_t = r(v^*_t)$ if $v_t \in V^*$ and $v_{t-1} \not\in V^*$. Finally, $\gamma = 1$, and the action is fixed to $a_0$, no matter what state, that is, $\pi(a|v) = a_0$ for all $v$.

Then for every $v$, with state transition probabilities $p(v'|v, a) = p(v'|v)$, we have:

$$G(v) = \sum_{V^* \cap V^*} r(v^*_t) \sum_{k=1}^{\infty} p^k(v^*_t|v) = \sum_{v^* \cap V^*} r(v^*_t) p^\infty(v^*_t|v).$$ (7)

where $p^k(v^*_t|v)$ is the $k$-step transition probability from $v$ to $v^*_t$, and $p^\infty(v^*_t|v)$ is the probability of eventually reaching $v^*_t$ from $v$.

As an example, consider a system that always evolves through three timesteps to reach an absorbing state, and at each timestep may be in one of three possible states. That is, the set of states are $V_0 \equiv \{v_0, v_1, v_2, v_3\}$, $V_1 \equiv \{v_11, v_12, v_13\}$
and $V_2 \equiv V^* \equiv \{v_1^*, v_2^*, v_3^*\}$. We have the following simple transition diagram: $v_{0i} \rightarrow v_{1j} \rightarrow v_{2k}^*$ for all $i, j, k$ and the above expression for total expected reward yields:

$$\sum_{v_0, v_1, v_2^*} r(V^*) p(V^*|V_1, a_0) p(V_1|V_0, a_0) p(V_0) = \mathbb{E}_{q_{e}}[r(V^*)],$$

(8)

where the expectation is taken with respect to the distribution $p(V^*|V_1, a_0) p(V_1|V_0, a_0) p(V_0)$. If we have a deterministic policy $\pi(a|v)$ (that sets each $v_i$ to a corresponding $a_i$), we have

$$\sum_{v_0, v_1, v_2^*} r(V^*) p(V^*|V_1, a_1 = \pi(V_1)) p(V_1|V_0, a_0) = \pi(V_0) p(V_0) = \mathbb{E}_{q_{e}}[r(V^*)],$$

(9)

Equations (8) and (9) resemble special cases of the $g$-formula, where structure of each state is simply represented by a single variable.

This special case illustrates that classical MDPs, despite allowing complicated state transition structure, have an important modeling disadvantage: they have difficulties handling confounding and other types of complex causal relationships within a state, and across states.

### C Graph preliminaries

Let capital letters $X$ denote random variables, and let lower case letters $x$ values of $X$. Sets of random variables are denoted $V$, and sets of values $v$. For a subset $A \subseteq V$, $V_A$ denotes the subset of values in $v$ of variables in $A$. Domains of $X$ and $X$ are denoted by $X$ and $X$, respectively.

Standard genealogic relations on graphs are as follows: parents, children, descendants, siblings and ancestors of $X$ in a graph $G$ are denoted by $pa_G(X), ch_G(X), dec_G(X), sis_G(X), an_G(X)$, respectively [Lauritzen 1996]. These relations are defined disjunctively for sets, e.g. $pa_G(X) \equiv \bigcup_{x \in X} pa_G(x)$. By convention, for any $X$, $an_G(X) \cap dec_G(X) \cap sis_G(X) = \{X\}$.

We will also define the set of strict parents as follows: $pa_G^0(X) = pa_G(X) \setminus X$. Given any vertex $V$ in an ADMG $G$, define the ordered Markov blanket of $V$ as $mb_G(V) \equiv (\text{dis}(V) \cup pa_G(\text{dis}(V))) \setminus V$. Given a graph $G$ with vertex set $V$, and $S \subseteq V$, define the induced subgraph $G_S$ to be a graph containing the vertex set $S$ and all edges in $G$ among elements in $S$.

### D The Nested Markov Factorization

It is recommended that the reader look up notation for graphs in Section C of the Appendix to follow this section.

#### D.1 Why do we need an alternative factorization?

A hidden variable CDAG $G(V \cup H, W)$ may be used to define a factorization on distributions $p(V|W)$ in terms of the CDAG as: $p(V|W) = \sum_H \prod_{V \in V \cup H} p(V|pa_G(V))$. However, inferences may be sensitive to assumptions made about the state spaces for the unobserved variables and the latent variable model may contain singularities at which asymptotics are irregular [Drton et al. 2009]. Additionally, such a model does not form a tractable search space: an arbitrary number of hidden variables and associated structures may be incorporated that are consistent with observed data distributions.

Alternatively, a factorization of the marginal distribution $p(V|W)$ can be defined directly on the latent projection CADMG $G(V, W)$. This nested Markov factorization, described in [Richardson et al. 2017] completely avoids modeling hidden variables, and leads to a regular likelihood in special cases [Evans and Richardson 2018]. It captures all equality constraints a hidden variable CDAG factorization imposes on the observed margin $p(V|W)$ [Spirtes et al. 2018]. In addition, $p(Y(a)|W)$ (an interventional distribution given a fixed context $W$) identified in a hidden variable causal model represented by $G(V \cup H, W)$ is always equal to a modified version of a nested factorization [Richardson et al. 2017] associated with $G(V, W)$, described here.

#### D.2 The nested Markov factorization

The nested Markov factorization of $p(V|W)$ with respect to a CADMG $G(V, W)$ links kernels, mappings derived from $p(V|W)$ and CADMGs derived from $G(V, W)$ via a fixing operation.

**Kernel:** A kernel $q_{V}(V|W)$ is a mapping from values in $W$ to normalized densities over $V$ [Lauritzen 1996]. A conditional distribution is a familiar example of a kernel, in that $\sum_{v \in V} q_{V}(v|w) = 1$. Conditioning and marginalization are defined in kernels in the usual way: For $A \subseteq V$, $q_{V}(A|W) \equiv \sum_{v \in A} q_{V}(v|W)$ and $q_{V}(\{v\} \setminus A|A \cup W) \equiv \frac{q_{V}(v|W)}{q_{V}(\{v\}|A \cup W)}$.

**Fixability and the fixing operator:** A variable $V \in V$ in a CADMG $G$ is fixable if $de_G(V) \cap de_G(V) = \emptyset$. In other words, $V$ is fixable if paths $V \leftrightarrow \ldots \leftrightarrow B$ and $V \rightarrow \ldots \rightarrow B$ do not both exist in $G$ for any $B \in V \setminus \{V\}$.

We define a fixing operator $\phi_{V}(G)$ for graphs, and a fixing operator $\phi_{V}(g; G)$ for kernels. Given a CADMG $G(V, W)$, with a fixable $V \in V$, $\phi_{V}(G(V, W))$ yields a new CADMG $G(V \setminus \{V\}, W \cup \{V\})$ obtained from $G(V, W)$ by moving $V$ from $V$ to $W$, and removing all edges with arrowheads into $V$. Given a kernel $q_{V}(V|W)$, and a CADMG $G(V, W)$, the operator $\phi_{V}(q_{V}(V|W), G(V, W))$ yields a new kernel:

$$q_{V \setminus \{V\}}(V \cup \{V\}) = \frac{q_{V}(V|W)}{q_{V}(V|mb_G(V))}$$

**Fixing sequences:** A sequence $(V_1, \ldots, V_k)$ is said to be valid in $G(V, W)$ if $V_1$ fixable in $G(V, W)$, $V_2$ is fixable in $\phi_{V_1}(G(V, W))$, and so on. If any two sequences $\sigma_1, \sigma_2$ for the same set $S \subseteq V$ are fixable in $G$, they lead to the same CADMG. The graph fixing operator can be extended to a set $S$:
Intrinsic sets: A set \( R \) reachable in \( G(V, W) \) is called intrinsic in \( G(V, W) \) if \( \phi_R \) contains a single district, \( R \) itself. The set of intrinsic sets in a CADMG \( G \) is denoted by \( \mathcal{I}(G) \).

Nested Markov factorization: A distribution \( p(V|W) \) is said to obey the nested Markov factorization with respect to the CADMG \( G(V, W) \) if there exists a set of kernels of the form \( \{ q_S|pa_S(S) : S \in \mathcal{I}(G) \} \) such that for every valid sequence \( \sigma_R \) for a reachable set \( R \) in \( G \), we have:

\[
\phi_{\sigma_R}(p(V|W); G(V, W)) = \prod_{D \in \mathcal{D}(\phi_R(G(V, W)))} q_D(D|pa^*_G(D))
\]

If a distribution obeys this factorization, then for any reachable \( R \), any two valid sequences on \( R \) applied to \( p(V|W) \) yield the same kernel \( \phi_R(R|V \setminus R) \). Hence, kernel fixing may be defined on sets, just as graph fixing. In this case, for every \( D \in \mathcal{I}(G) \), \( q_D(D|pa^*_G(D)) = \phi_{\mathcal{D}(\phi_R(G(V, W)))}(p(V|W); G(V, W)) \).

The district factorization or Tian factorization of \( p(V|W) \) results from the nested factorization:

\[
p(V|W) = \prod_{D \in \mathcal{D}(\phi_R(G(V, W)))} q_D(D|pa^*_G(D)) = \prod_{D \in \mathcal{D}(\phi_R(G(V, W)))} \left( \prod_{D \in \mathcal{D}(\phi_R(G(V, W)))} p(D | \text{pre}_{\prec}(D)) \right),
\]

where \( \text{pre}_{\prec}(D) \) is the set of predecessors of \( D \) according to a topological total ordering \( \prec \). Each factor \( \prod_{D \in \mathcal{D}(\phi_R(G(V, W)))} p(D | \text{pre}_{\prec}(D)) \) is only a function of \( D \cup pa_G(D) \) under the nested factorization.

An important result in [Richardson et al. 2017] states that if \( p(V \cup H|W) \) obeys the factorization for a CDAG \( G(V \cup H, W) \), then \( p(V|W) \) obeys the nested factorization for the latent projection CADMG \( G(V, W) \).

D.3 Identification

Not every interventional distribution \( p(Y(a)) \) is identified in a hidden variable causal model. However, every \( p(Y(a)|W) \) identified from \( p(V|W) \) can be expressed as a modified nested factorization as follows:

\[
p(Y(a)|W) = \sum_{Y^* \setminus V \in \mathcal{D}(\phi_V(G(V, W)))} \prod_{Y \setminus V \in \mathcal{D}(\phi_V(G(V, W)))} \phi_{\mathcal{D}(\phi_V(G(V, W)))}(p(V|W); G(V, W)|_{A=a})
\]

where \( Y^* \equiv \cup \mathcal{D}(\phi_V(G(V, W))) \). That is, \( p(Y(a)|W) \) is only identified if it can be expressed as a factorization, where every piece corresponds to a kernel associated with a set intrinsic in \( G(V, W) \). Moreover, no piece in this factorization contains elements of \( A \) as random variables.

D.4 Parameterization of Binary Nested Markov Models

The familiar DAG model for binary variables is parameterized by factors of the form \( P(X_i = 0 | pa(X_i), \text{corresponding to each node } X_i \text{ in the model. Similarly, the binary nested Markov model can be parameterized by associating with each intrinsic set, 'heads' and 'tails' for each } B \in \mathcal{I}(G), \text{comparable to } X_i \text{ and } pa(X_i) \) respectively [Richardson et al. 2012] [Shpitser et al. 2012].

Heads and tails of intrinsic sets: For an intrinsic set \( B \in \mathcal{I}(G) \) of a CADMG \( G(V, W) \), the recursive head is defined as \( r_h(B) \equiv \{ x | x \in B; ch_{\phi_B}(x) = \emptyset \} \). Let \( \mathcal{R}_h(G) \equiv \{ r_h(B) | B \in \mathcal{I}(G) \} \). The tail associated with a recursive head \( H \) of an intrinsic set \( B \) in a CADMG \( G \) is given by: \( \text{tail}(H) \equiv \{ B \setminus H \} \cup pa_G(B) \).

Partition of arbitrary sets: Any arbitrary subset of \( V \) in a CADMG \( G(V, W) \) can be partitioned into elements of \( \mathcal{R}_h(G) \). Let \( \prec (G) \) be a partial order on heads of intrinsic sets of \( G \) such that \( H_1 \prec (G) H_2 \) if \( H_1 = r_h(B_1), H_2 = r_h(B_2), B_1, B_2 \in \mathcal{I}(G), B_1 \subseteq B_2 \).

For a set of heads \( H \), let \( \text{max}_{\prec}(G) \) be the subset of \( H \) containing heads maximal in \( H \) under \( \prec (G) \). For any \( B \subseteq V \) of nodes in a CADMG \( G(V, W) \), we define:

\[
\mathcal{Y}_G(B) \equiv \text{max}_{\prec(G)}(\mathcal{R}_h(G) \cap \mathcal{P}(B))
\]

\[
\rho_G(B) \equiv B \setminus \bigcup_{H \in \mathcal{Y}_G(B)} \rho_G^{(k)}(B) \equiv \rho_G^{(k)}(\ldots \rho_G^{(k)}(B) \ldots),
\]

where \( \mathcal{P}(B) \) is the power set of \( B \) and \( \rho_G^{(0)}(B) \equiv B \). We partition \( B \) into recursive heads of \( G \) as follows:

\[
\{B \} \equiv \bigcup_{k \geq 0} \mathcal{Y}_G\left( \rho_G^{(k)}(B) \right)
\]

Binary Parameterization: Multivariate binary distributions which obey the nested factorization with respect to a CADMG \( G(V, W) \) may be parameterized by a set of functions:

\[
\Omega_G \equiv \{ q_B(X_H = 0 | x_{\text{tail}(H)} | H = r_h(B), B \in \mathcal{I}(G)) \}.
\]
Intuitively, a parameter $q_B (X_H = 0 | x_{tail(H)})$ is the probability that the variable set $X_H$ assumes values 0 in a kernel obtained from $p(x_V)$ by fixing $X_{V \setminus B}$, and conditioning on $X_{tail(H)}$. To simplify notation, we will denote the parameter $q_B (X_H = 0 | x_{tail(H)})$ as $\theta_H (x_{tail(H)})$.

Let $\nu : V \cup W \rightarrow \{0, 1\}$ be an assignment of values to the variables indexed by $V \cup W$. Define $\nu (T)$ to be the values assigned to variables indexed by a subset $T \subseteq V \cup W$. Let $\nu^{-1}(0) = \{ v \in V, \nu(v) = 0 \}$

Suppose that $D_1 \cup D_2 \cup D_3 \cdots \cup D_k = V_G$ and each pair $D_i$ and $D_j$, $i \neq j$ are disconnected in $G$. A distribution $P(X_V \setminus X_W)$ is said to be parameterized by the set $D_G$, for CADMG $G$ if:

$$p(X_V = \nu(V) \mid X_W = \nu(W)) = \frac{k_i}{\prod_{i} (\nu^{-1}(0) \cap D_i)} \prod_{\theta_H (x_{tail(H)}) = \nu(\text{tail}(H))}$$

where the empty product is defined to be 1, and $[[B]]_p$ is a partition of nodes in $B$ given earlier.

Thus we can factorize the parametrization into districts; note that this does not imply independence between districts, since the tail of a head in one district may contain vertices in another. Additionally, there is in general no partial order on heads such that each tail of a head is contained within the earlier heads.

Implications for Statistical Inference For Latent Variable PDSEMs: If a PDSEM is fully observed, causal inference may be performed by obtaining maximum likelihood estimates $\hat{\eta}$ of all parameters, and evaluating the g-formula functions using Monte-Carlo sampling using learned distributions of the form $p(V \mid p_a g(V); \eta_V)$. This method is computationally efficient as long as the initial DAG and transition CDAGs in a PDSEM are sufficiently sparse. Indeed, both our simulations and our data application were based on this approach.

However, an analogous approach does not work for nested Markov parameterizations of the marginal PDSEM representing a PDSEM with hidden variables. This is because the above Möbius parameterization is ill-suited for drawing samples. Instead, existing approaches to sampling from a nested Markov discrete likelihood involve first converting the likelihood expressed in terms of the Möbius parameters to one expressed as a joint distribution $p(V)$ (from which it is easy to generate samples for a discrete sample space of $V$). Importantly, such a conversion leads to an intractable object that requires storage and running time exponential in $|V|$. This holds even if the underlying model dimension of the nested Markov model is small. The situation is radically different from that of DAG models, where a small model dimension directly leads to a computationally efficient sampling scheme.

While there exist promising approaches, based on the nested Markov generalization of the variable elimination algorithm [Shpitser et al. 2011], in general the problem remains open.

E Proofs

**Lemma 1** For any $A \subseteq V$, $p(V \setminus A | a) | W$ is identified from $p(V | W)$ in the functional model of $G(V \setminus W)$ as

$$p(V \setminus A | a) | W = \prod_{V \setminus A} p(V | p_a g(V)) | A = a$$

**Proof:** This follows by a simple generalization of Proposition 17 in [Richardson and Robins 2013].

**Lemma 2** Given a fully observed PDSEM, $p_\infty(V(a))$ is identified from $p_\infty(V)$ as:

$$p_1(V_1(a_1)) \equiv \prod_{a_1, \nu(V_1 | A_1)} p(V | pa_{g_1}(V)) | A = a_1$$

$$p_1(V_1(a_1), V_2(a_2)) \equiv \prod_{a_1, a_2, \nu(V_1 | A_1), \nu(V_2 | A_2)} p(V | pa_{g_2}(V)) | A = a_2 \quad (11)$$

**Proof:** This follows from the factorization of $p_\infty(V(a))$ into elements of the form $p(V_1(a_1))$, and $p(V_1(a_1), V_2(a_2))$. The fact that $G_1 \cup G_2$ define causal models under standard structural equation semantics, and Lemma 1

**Lemma 3** Given the latent variable PDSEM, $p_\infty(V(a))$ is identified from $p_\infty(V)$ if and only if

$$p(V_i \setminus A_i | a_i) = \prod_{D \subseteq D_i \cup G_i} q_D(D | p_a g_i(D))$$

with each kernel $q_D(D | p_a g_i(D))$ above is identified from $p_1(V_i)$, and for each $(i,j) \notin T$,

$$p(V_i \setminus A_i | a_i) | W_i = \prod_{D \subseteq D_i \cup G_i} q_D(D | p_a g_i(D))$$

with each kernel $q_D(D | p_a g_i(D))$ above is identified from $p_1(V_i | W_i)$. **Proof:** This follows from the factorization of the marginal distribution $p_\infty(V(a))$ into elements of the form $p(V_1(a_1))$, and $p(V_1(a_1), V_2(a_2))$. The fact that $G_1 \cup G_2$ are latent projections of hidden variable causal models under standard structural equation semantics, and Theorem 60 in [Richardson et al. 2017].

F Computation Details

The septoplasty data application presented in Section 5 was computed on a Lenovo X1 Carbon with an Intel i7 1.8 GHz processor and 16 GB of RAM. Computation for each scenario (generating from the model without interventions, attending performing the whole surgery, and trainee performing the whole surgery) took between 1.5 to 2 hours.

References

Narges Ahmidi, Piyush Poddar, Jonathan D Jones, S Swareep Vedula, Lisa Ishii, Gregory D Hager, and Masaru Ishii. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *International journal of computer assisted radiology and surgery*, 10(6):981–991, 2015.

C. F. Belanger, C. H. Hennekens, B. Rosner, and F. E. Speizer. The nurses’ health study. *The American Journal of Nursing*, 78(6):1039–1040, June 1978. ISSN 0002-936X.

B. Chakraborty and EE. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. New York: Springer-Verlag, 2013.

Mathias Drton et al. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012, 2009.

Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 2018. (to appear).
Nicholas Fettman, Thomas Sanford, and Raj Sindwani. Surgical management of the deviated septum: techniques in septoplasty. Otolaryngologic Clinics of North America, 42(2):241–252, 2009.

Miguel Angel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. Epidemiology, pages 561–570, 2000.

Sebastian Hohna, Tracy A. Heath, Bastien Boussau, Michael J. Landis, Fredrik Ronquist, and John P. Huelsenbeck. Probabilistic graphical model representation in phylogenetics. Systematic Biology, 63:753–771, 2014.

Timothy D Jackson, Jeffrey J Wannares, R Todd Lancaster, David W Rattner, and Matthew M Hutter. Does speed matter? the impact of operative time on outcome in laparoscopic surgery. Surgical endoscopy, 25(7):2288–2295, 2011.

Stefan L. Lauritzen. Graphical Models. Oxford, U.K.: Clarendon, 1996.

Peter Lee, S Regenbogen, and AA Gawande. How many surgical procedures will americans experience in an average lifetime?: Evidence from three states. In 55th Annual Meeting, 2008.

Stan Liebowitz and Stephen Margolis. Path dependence. Encyclopedia of Law and Economics, 2002.

Michael L Littman. A tutorial on partially observable markov decision processes. Journal of Mathematical Psychology, 53(3):119–125, 2009.

Daniel Malinsky and Peter Spirtes. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. In Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery, pages 23–47, August 2018.

Daniel Malinsky and Peter Spirtes. Learning the Structure of a Non-stationary Vector Autoregression. Proceedings of machine learning research, 89:2986–2994, April 2019. ISSN 2640-3498.

Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press, September 2012. ISBN 978-0-262-30432-0.

Susan A Murphy. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):331–355, 2003.

Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000. ISBN 0-521-77362-8.

Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2 edition, 2009. ISBN 978-0521895606.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal Inference on Time Series using Restricted Structural Equation Models. Advances in Neural Information Processing Systems 26, pages 154–162, 2013.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of the IEEE, volume 77, pages 257–286, 1989.

Aksharananda Rambachan, Lauren M Mioton, Sujata Saha, Neil Fine, and John YS Kim. The impact of surgical duration on plastic surgery outcomes. European Journal of Plastic Surgery, 36(11):707–714, 2013.

Thomas S. Richardson and Jamie M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. preprint: http://www.csss.washington.edu/Papers/wp128.pdf 2013.

Thomas S. Richardson, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. In Twenty Eighth Conference on Uncertainty in Artificial Intelligence (UAI-12), 2012.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017. Working paper.

James M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. Mathematical Modeling, 7:1393–1512, 1986.

James M. Robins. Marginal structural models. In American Statistical Association, Section on Bayesian Statistical Science, pages 1–10, 1997.

James M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In Statistical Models in Epidemiology: The Environment and Clinical Trials. NY: Springer-Verlag, 1999.

Ilya Shpitser, Thomas S. Richardson, and James M. Robins. An efficient algorithm for computing interventional distributions in latent variable causal models. In Uncertainty in Artificial Intelligence, volume 27. AUAI Press, 2011.

Ilya Shpitser, Thomas S. Richardson, James M. Robins, and Robin J. Evans. Parameter and structure learning in nested Markov models. In Causal Structure Learning Workshop of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-12), 2012.

Ilya Shpitser, Robin J. Evans, and Thomas S. Richardson. Acyclic linear sems obey the nested markov property. In Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18), 2018.

C. Spearman. Human Ability. Macmillian, London, 1950.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. MIT Press, October 2018. ISBN 978-0-262-35270-3.

Bobby A Tajudeen and David W Kennedy. Thirty years of endoscopic sinus surgery: What have we learned? World journal of otorhinolaryngology-head and neck surgery, 3(2):115–121, 2017.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02), volume 18, pages 519–527. AUAI Press, Corvallis, Oregon, 2002.