Path Dependent Structural Equation Models

Ranjani Srinivasan[1] **Jaron J. R. Lee**[2] **Rohit Bhattacharya**[2] **Ilya Shpitser**[2]

[1]Electrical and Computer Engineering, Johns Hopkins University, 3400 N Charles St, Baltimore, MD USA, 21218
[2]Computer Science, Johns Hopkins University, 3400 N Charles St, Baltimore, MD USA, 21218

Abstract

Causal analyses of longitudinal data generally assume that the qualitative causal structure relating variables remains invariant over time. 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. Tasks such as discovering long term health risk factors [Belanger et al., 1978], object trajectory tracking [Richards, 2005], and speech recognition [Rabiner, 1989] all require modeling temporal evolution of relationships among a set of variables. Many models for longitudinal 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, or find latent structures that best explain observations, and deal fundamentally with *associative* relationships. However, testing empirical hypotheses or providing decision support often requires *causal* modeling. Causal models based on graphs [Pearl, 2009, Richardson and Robins, 2013] have become increasingly popular in part because of their ability to display complex relationships in multivariate systems in an intuitive visual way. These models have been extended to *dynamic causal Bayesian networks* [Blondel et al., 2017] that can model causal relationships in temporal processes that evolve in discrete time. However, these models have generally been used in settings with causal structure that remains invariant over time. For example, analysis of the impact of anti-retroviral therapy on HIV infection progression assumed the same variables relevant for the patient health and the same causal relationships linking them at each time point in the study [Hernán et al., 2000]. Changes tracked over time (such as HIV developing resistance to the current drug) are thus *quantitative*, with the underlying causal structure remaining unchanged over time. However, many systems undergo *qualitative* changes as well, where observability, relevance, and causal relationships of variables vary over time.

Consider the task of modeling surgical procedures to make informed decisions on resident surgeon training. Surgeries are often divided into discrete stages, each with an intermediate goal [Ahmidi et al., 2015]. Each stage is associated with a distinct set of variables and relationships among them that may not be shared across stages. For instance, stitching together a previously made incision is a routine task requiring few tools that may be executed by a surgical robot, while reconstructing cartilage is a skill-intensive task requiring multiple tools, high surgical skill and manual dexterity. Another feature of surgeries is that procedures performed at a particular stage can go wrong, forcing surgeons to "double back" to correct mistakes or deal with complications. Surgeon experience often determines how likely it is that previous stages of the surgery need to be revisited.

The goal of causal inference in this setting is to help assign surgeons to perform different stages of the surgery while navigating the tradeoff between the need to train resident surgeons on the one hand, and operating costs and patient safety on the other. Addressing this tradeoff entails using ret-

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:161–171.

rospective data to estimate outcomes of surgery trajectories that *differ from those actually observed* due to counterfactually different choices of surgeon assignment in past stages of the surgery. Following the convention in the economics literature, we call the phenomenon where the evolution of a system changes in response to counterfactually different past choices *path dependence* [Liebowtiz and Margolis, 2002]. Other examples where path dependence may naturally arise include life course studies examining economic disparities in society or patient outcomes in hospitals using Electronic Health Record (EHR) data.

Our contributions to the literature are as follows. We introduce the *path-dependent structural equation model (PDSEM)* for causal systems that exhibit qualitative changes over time, observed or unobserved confounding, *and* path-dependence on counterfactual choices in the past. PDSEMs generalize causal dynamic Bayesian networks by allowing complex and looping stage transitions between distinct yet tractable causal models, and generalize Markov decision processes used in reinforcement learning [Sutton and Barto, 2018, Zhang and Bareinboim, 2016] by representing each state as a graphical causal model that allows confounding between actions and outcomes. We give a complete identification theory for our model. In particular, in the special case where the PDSEM is first order Markov, all identification queries may be decomposed into queries pertaining to observed transition probabilities between states, a generalization of results for causal DBNs in [Blondel et al., 2017]. Finally, we show how statistical inference may be performed by a combination of plug-in estimation and Monte Carlo sampling, generalizing similar schemes developed for longitudinal causal models [Westreich et al., 2012].

2 BACKGROUND

We review preliminaries on graphical causal modeling, before discussing extensions that allow path-dependence.

2.1 STATISTICAL AND CAUSAL DAG MODELS

The statistical model of a directed acyclic graph (DAG) $\mathcal{G}(\mathbf{V})$ with a vertex set $\mathbf{V} \equiv \{V_1, \ldots, V_k\}$, also called a *Bayesian network*, is the set of distributions that Markov factorize with respect to the DAG as $p(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i))$ where $\text{pa}_{\mathcal{G}}(V_i)$ are parents of V_i in $\mathcal{G}$.

Causal models of a DAG are also sets of distributions but on *counterfactual* random variables. Each variable V_i in a causal model is determined from values of its parents $\text{pa}_{\mathcal{G}}(V_i)$ and an exogenous noise variable ϵ_i via an invariant causal mechanism called a *structural equation* $f_i(\text{pa}_{\mathcal{G}}(V_i), \epsilon_i)$. Causal models allow counterfactual intervention operations, denoted by the $\text{do}(\mathbf{a})$ operator in [Pearl, 2009]. Such operations replace each structural equation $f_i(\text{pa}_{\mathcal{G}}(V_i), \epsilon_i)$ for $V_i \in \mathbf{A} \subset \mathbf{V}$ by one that sets V_i to a constant value in $\mathbf{a}$ corresponding to V_i. The joint distribution of variables in $\mathbf{Y} \equiv \mathbf{V} \setminus \mathbf{A}$ after the intervention $\text{do}(\mathbf{a})$ was performed is denoted by $p(\mathbf{Y} \mid \text{do}(\mathbf{a}))$, equivalently written as $p(\{V_i(\mathbf{a}) : V_i \in \mathbf{Y}\})$, or $p(\mathbf{Y}(\mathbf{a}))$, where $V_i(\mathbf{a})$ is a counterfactual random variable or a potential outcome.[1]

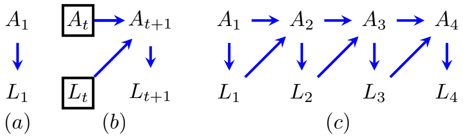

Figure 1: (a) Prior network DAG $\mathcal{G}_0$, representing the state of the dynamic Bayesian network at time $t = 0$. (b) A conditional DAG $\mathcal{G}_{t,t+1}$ representing the transitions in a dynamic Bayesian network. (c) A dynamic Bayesian network model unrolled to four time steps.

A popular causal model called the *non-parametric structural equation model with independent errors (NPSEM-IE)* [Pearl, 2009] assumes, aside from the structural equations for each variable being functions of their parents in the DAG $\mathcal{G}(\mathbf{V})$, that the joint distribution of all exogenous terms are marginally independent: $p(\epsilon_1, \epsilon_2, \ldots) = \prod_{V_i \in \mathbf{V}} p(\epsilon_i)$. The NPSEM-IE implies the DAG factorization of $p(\mathbf{V})$ with respect to $\mathcal{G}(\mathbf{V})$, and a truncated DAG factorization known as the *g-formula*:

$$p(\mathbf{Y}(\mathbf{a})) = \prod_{V_i \in \mathbf{Y}} p(V_i \mid \text{pa}_{\mathcal{G}}(V_i))|_{\mathbf{A}=\mathbf{a}} \tag{1}$$

for every $\mathbf{A} \subseteq \mathbf{V}$, and $\mathbf{Y} = \mathbf{V} \setminus \mathbf{A}$.

2.2 GRAPHICAL MODELS IN DISCRETE TIME

While Bayesian networks lend themselves well to the modeling of static data, data that changes over time requires more sophisticated models. A generalization of the Bayesian network model for discrete time temporal systems is the *dynamic Bayesian network (DBN)* model [Murphy, 2012].

A DBN is specified by a pair of DAGs, and a corresponding pair of factorized distributions. The *prior network* $\mathcal{G}_1$ and its corresponding distribution $p(\mathbf{V}_1) = \prod_{V_i \in \mathbf{V}_1} p(V_i \mid \text{pa}_{\mathcal{G}_1}(V_i))$ represent the state of the system at the first time step. The *transition network* $\mathcal{G}_{t,t+1}$ is a *conditional DAG (CDAG)* with random vertices $\mathbf{V}_{t+1}$ representing

[1] Our use of the word *counterfactual* here follows standard usage in the statistics and public health literature, see e.g. [Hernan and Robins, 2020], and represents the fact that $V_i(\mathbf{a})$ represents the outcome V_i if variables $\mathbf{A}$ were altered *possibly contrary to fact* to attain values $\mathbf{a}$. Another sense of this word refers to a change of an actually occurring prior event, see e.g. the discussion of abductive inference in twin network models in [Pearl, 2009]. The models we develop in this paper also allow us to consider counterfactually different temporal evolution in the latter sense of the word.

variables at time point $t+1$, and fixed vertices $\mathbf{V}_t$ representing context in the previous time point t. We will describe such conditional graphs by a shorthand "$\mathcal{G}_{t+1,t}$ on $\mathbf{V}_{t+1}$ given $\mathbf{V}_t$". In this conditional DAG no arrowheads into vertices in $\mathbf{V}_t$ are allowed. The corresponding conditional distribution $p(\mathbf{V}_{t+1} \mid \mathbf{V}_t)$ represents the way variables at point $t+1$ depend on each other, and on variables at the prior time point t (and on no other prior variables, such as those at time point $t-1$). This dependence leads to a *first-order Markov* DBN. This distribution factorizes with respect to the CDAG $\mathcal{G}_{t,t+1}$ as follows: $p(\mathbf{V}_{t+1} \mid \mathbf{V}_t) = \prod_{V_i \in \mathbf{V}_{t+1}} p(V_i \mid \text{pa}_{\mathcal{G}_{t,t+1}}(V_i))$.

The joint distribution for the DBN system over a finite number of discrete time steps T is given by the product of the prior network distribution, and the transition conditional probability distributions for a set of time steps, as follows:

$$\prod_{V \in \mathbf{V}_1} p(V | \text{pa}_{\mathcal{G}_1}(V)) \cdot \prod_{t=1}^{T-1} \prod_{V \in \mathbf{V}_{t+1}} p(V | \text{pa}_{\mathcal{G}_{t,t+1}}(V)) \quad (2)$$

A simple DBN is shown in Figure 1, where the prior network (1(a)) contains two variables A and L, and the transition network (1(b)) shows connections among the state variables in the prior state at time t and the subsequent state at time $t+1$. We represent fixed vertices in a transition network via squares. Figure 1(c) shows the DBN implied by these prior and transition networks unrolled to 4 time steps.

DBNs can be naturally extended to represent causal models by assuming that both prior and transition networks are causal DAGs. In other words, we assume values of every variable V_i in both the prior and the transition network is determined, via a structural equation $f_i(.)$, in terms of its observed parents $\text{pa}_{\mathcal{G}_1}(V_i)$ (or $\text{pa}_{\mathcal{G}_{t,t+1}}(V_i)$) and an exogenous noise term ϵ_i. If we further assume that all exogenous noise variables are marginally independent, we arrive at a DBN version of the NPSEM-IE, where in addition to the g-formula (1) holding for the prior network, the *conditional g-formula* holds for the transition network:

$$p(\mathbf{Y}_{t+1}(\mathbf{a}) | \mathbf{V}_t) = \prod_{V_i \in \mathbf{Y}_{t+1}} p(V_i | \text{pa}_{\mathcal{G}}(V_i))|_{\mathbf{A}=\mathbf{a}}, \quad (3)$$

for any $\mathbf{A} \subseteq \mathbf{V}_{t+1}$, and $\mathbf{Y}_{t+1} = \mathbf{V}_{t+1} \setminus \mathbf{A}$. Thus, a causal DBN "unrolled" to a set of time points $1, \ldots, T$ yields a standard causal DAG model with vertices $\mathbf{V}_{1:T} \equiv \mathbf{V}_1 \cup \mathbf{V}_2 \cup \ldots \cup \mathbf{V}_T$. For an intervention that sets $\mathbf{A} \subseteq \mathbf{V}_{1:T}$ to constant values $\mathbf{a}$, the interventional distribution $p(\mathbf{Y}_{1:T}(\mathbf{a}))$, where $\mathbf{Y}_{1:T} = \mathbf{V}_{1:T} \setminus \mathbf{A}$, is identified by:

$$\left. \prod_{V \in \mathbf{V}_1 \setminus \mathbf{A}} p(V | \text{pa}_{\mathcal{G}_1}(V)) \prod_{t=1}^{T-1} \prod_{V \in \mathbf{V}_{t+1} \setminus \mathbf{A}} p(V | \text{pa}_{\mathcal{G}_{t,t+1}}(V)) \right|_{\mathbf{A}=\mathbf{a}} \quad (4)$$

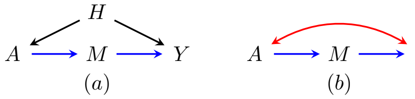

Figure 2: (a) A hidden variable DAG, and (b) its latent projection ADMG.

Causal DBNs have been considered in prior work. [Peters et al., 2013] illustrated how structural equations can be used in the context of time series data, addressing issues of identifiability. [Malinsky and Spirtes, 2018, 2019, Mogensen et al., 2018] presented structure learning algorithms for causal dynamic networks and applied them to macroeconomic data. [Blondel et al., 2017] developed an identification algorithm and transportability results for dynamic causal networks. The first-order Markov assumption in DBN models may be relaxed to a kth-order Markov assumption, where the model at any time step depends on variables in at most k prior time steps, a generalization we describe in Appendix C.2.

2.3 HIDDEN VARIABLE CAUSAL MODELS

The g-formula (1) provides an elegant link between observed data and counterfactual distributions in causal models where all relevant variables are observed. Causal models that arise in practice, however, contain hidden variables. Representing such models using a DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ where $\mathbf{V}$ and $\mathbf{H}$ correspond to observed and hidden variables, respectively, is not very helpful, since applying (1) to $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ results in an expression that involves unobserved variables $\mathbf{H}$. A popular alternative is to represent a class of hidden variable DAGs $\mathcal{G}_i(\mathbf{V} \cup \mathbf{H}_i)$ by a single *acyclic directed mixed graph* ADMG $\mathcal{G}(\mathbf{V})$ that contains directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges and no directed cycles via the *latent projection* operation [Verma and Pearl, 1990] (see Section B of the Appendix). The latent projection ADMG $\mathcal{G}(\mathbf{V})$ captures relationships between observed variables $\mathbf{V}$ implied by the factorization of $p(\mathbf{V} \cup \mathbf{H})$ with respect to $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ via the *nested Markov factorization* of $p(\mathbf{V})$ with respect to $\mathcal{G}(\mathbf{V})$ [Richardson et al., 2017].

In particular, just as identification in DAGs may be viewed in terms of a modified DAG factorization (1), identification in a hidden variable DAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H})$ may be viewed in terms of a modified nested factorization of $\mathcal{G}(\mathbf{V})$. The nested Markov factorization of $p(\mathbf{V})$ with respect to $\mathcal{G}(\mathbf{V})$ is defined in terms of *Markov kernels* of the form $q_{\mathbf{D}}(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})$, where set $\mathbf{D} \subseteq \mathbf{V}$ is *intrinsic* in $\mathcal{G}(\mathbf{V})$. Kernels $q_{\mathbf{D}}(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})$ are objects that resemble conditional densities $p(V_i \mid \text{pa}_{\mathcal{G}}(V_i))$ that arise in the Markov factorization for a DAG, in the sense that they are non-negative and normalize to 1 for every value of $\text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D}$. Kernels making up the nested Markov factorization are all functionals of $p(\mathbf{V})$. A set $\mathbf{S}$ is intrinsic in

$\mathcal{G}(\mathbf{V})$ if $p(\mathbf{S}|\text{do}(\text{pa}(\mathbf{S}) \setminus \mathbf{S}))$ is identified.

The nested Markov factorization asserts that the observed margin $p(\mathbf{V})$ can be expressed as a product $\prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}(\mathbf{V}))} q_{\mathbf{D}}(\mathbf{D} \mid \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})$ of kernels where $\mathcal{D}(\mathcal{G}(\mathbf{V}))$ is the set of bidirected connected components, called *districts*, in $\mathcal{G}(\mathbf{V})$. The factorization implies certain other kernels associated with *reachable sets* may be expressed as similar products of intrinsic kernels. Finally, the modified form of the factorization may be used to express *any* interventional distribution identified from $p(\mathbf{V})$.

Given a latent projection ADMG $\mathcal{G}(\mathbf{V})$ representing a hidden variable causal model, and any disjoint subsets $\mathbf{Y}, \mathbf{A}$ of $\mathbf{V}$, let $\mathbf{Y}^*$ be the set of ancestors of $\mathbf{Y}$ in $\mathcal{G}(\mathbf{V})$ via directed paths that do not pass through $\mathbf{A}$, and let $\mathcal{G}_{\mathbf{Y}^*}$ be the *induced subgraph* of $\mathcal{G}(\mathbf{V})$ containing only vertices in $\mathbf{Y}^*$ and edges among these vertices. [Shpitser and Pearl, 2006, Richardson et al., 2017] showed that any interventional distribution $p(\mathbf{Y}(\mathbf{a}))$ is identified from $p(\mathbf{V})$ given $\mathcal{G}(\mathbf{V})$ if and only if every bidirected connected component in $\mathcal{G}_{\mathbf{Y}^*}$ is intrinsic. Moreover, if $p(\mathbf{Y}(\mathbf{a}))$ is identified, it is given by the following margin of the modified nested Markov factorization, made up of the appropriate kernels:

$$p(\mathbf{Y}(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus (\mathbf{Y} \cup \mathbf{A})} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} q_{\mathbf{D}}(\mathbf{D} | \text{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})|_{\mathbf{A}=\mathbf{a}}. \quad (5)$$

As a simple example, consider the hidden variable DAG in Fig. 2(a). Its latent projection ADMG in Fig. 2(b), called the *front-door graph*, has intrinsic sets $\{A\}, \{M\}, \{A, Y\}, \{Y\}$, with the corresponding kernels: $q_A(A) \equiv p(A)$, $q_M(M|A) \equiv p(M|A)$, $q_{A,Y}(A, Y|M) \equiv p(Y|M, A)p(A)$, and $q_Y(Y \mid M) \equiv \sum_A p(Y|M, A)p(A)$.

By the nested Markov factorization, the observed margin $p(A, M, Y)$ factorizes as $q_{A,Y}(A, Y|M)q_M(M|A)$. In addition, certain other distributions also factorize. For example, the margin $p(A, M)$ is equal to $q_A(A)q_M(M|A)$. Further, $p(Y(a))$ is identified from $p(A, M, Y)$ and equal to $\sum_M q_Y(Y|M)q_M(M|a) = \sum_M \left(\sum_{A'} p(Y|M, A')p(A')\right) p(M|a)$, which is the *front-door formula* [Pearl, 1995]. See Section B of the Appendix for details on the nested Markov factorization, reachable and intrinsic sets, and identification theory in ADMGs.

3 IDENTIFICATION IN A CAUSAL DBN WITH HIDDEN VARIABLES

[Blondel et al., 2017] showed how identification in hidden variable causal DBNs may be decomposed into a set of independent problems, pertaining to conditional state transition distributions. We reformulate and generalize these results using the language of nested Markov models to facilitate identification theory and statistical inference in PDSEMs. We start with an assumption that allows us to view the marginal version of a DBN, defined only on observed variables, as a first-order Markov DBN.

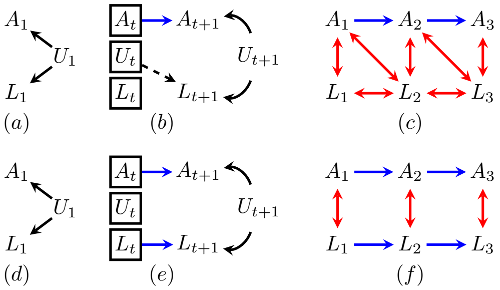

Figure 3: (a),(d) Prior network hidden variable DAGs $\mathcal{G}_0$, representing the state at time $t = 0$. (b),(e) Conditional hidden variable DAGs $\mathcal{G}_{t,t+1}$ representing the transitions in the network, with (e) leading to a first-order Markov model, and (b) leading to higher order dependences to unobserved hidden variables U_t linking multiple time points. (c),(f) Latent projection ADMGs of the unrolled hidden variable DBNs to three time steps.

Assumption 1 *Transition network $\mathcal{G}_{t+1,t}$ only depends on fixed variables in the previous time step t that are observed.*

If $\mathcal{G}_{t+1,t}$ depends on fixed variables that are hidden, the resulting DBN may result in observed variables in step $t + 1$ depending on observed variables earlier than t even if observed variables in t are conditioned on, resulting in a model that is not first order Markov.

For example, consider the DBN specified by prior and transition networks in Fig. 3 (a) and (b). Because the variable L_{t+1} depends on U_t, which is unobserved, and U_t influences L_t, "unrolling" this network, and taking the latent projection yields an ADMG shown in Fig. 3 (c), where L_3 ends up being dependent on L_1, even after conditioning on L_2, A_2 (due to the "explaining away" phenomenon arising when a shared effect L_2 of two variables U_2 and U_1 is conditioned on). On the other hand, the DBN specified by prior and transition networks in Fig. 3 (d) and (e) does not suffer from this issue, as the transition network only depends on observed variables L_t, A_t, yielding a latent projection of the "unrolled" model shown in Fig. 3 (f), which factorizes into time step specific conditional distributions: $p(A_1, L_1)p(A_2, L_2|A_1, L_1)p(A_3, L_3|A_2, L_2)$.

In general, given a hidden variable prior network $\mathcal{G}_1$ on $\mathbf{V}_1, \mathbf{H}_1$, and transition network $\mathcal{G}_{t+1,t}$ on $\mathbf{V}_{t+1}, \mathbf{H}_{t+1}$ given $\mathbf{V}_t$, the hidden variable DBN may be represented by latent projections of the prior and transition networks: an ADMG $\mathcal{G}_1$ on $\mathbf{V}_1$, and a *conditional ADMG (CADMG)* $\mathcal{G}_{t+1,t}$ on $\mathbf{V}_{t+1}$ given $\mathbf{V}_t$, and the corresponding marginal distributions $p(\mathbf{V}_1)$ and $p(\mathbf{V}_{t+1,t}|\mathbf{V}_t)$. The "unrolled" version of the factorization of this model is: $p(\mathbf{V}_1) \prod_{t=1}^{T} p(\mathbf{V}_{t+1,t}|\mathbf{V}_t)$, where each term nested

Markov factorizes with respect to either $\mathcal{G}_1$ or $\mathcal{G}_{t+1,t}$ by results in [Richardson et al., 2017]. [2]

If the underlying DAGs correspond to causal models, the hidden variable DBN yields identification theory where modified nested factorization (5) is applied at every time point, just as (1) was applied at every point in a fully observed causal DBN to yield (4). Given a fixed set of time points $1, \ldots, T$, vertices $\mathbf{V}_{1:T} \equiv \mathbf{V}_1 \cup \mathbf{V}_2 \cup \ldots \cup \mathbf{V}_T$, and disjoint subsets $\mathbf{A}, \mathbf{Y} \subseteq \mathbf{V}_{1:T}$, we have the following generalization of results in [Blondel et al., 2017]:

Lemma 1 *Under Assumption 1 , $p(\mathbf{Y}(\mathbf{a}))$ is identified from a hidden variable causal DBN model represented by latent projections $\mathcal{G}_1$ on $\mathbf{V}_1$ and $\mathcal{G}_{t+1,t}$ on $\mathbf{V}_{t+1}$ given $\mathbf{V}_t$ if and only if every bidirected connected component in $\mathcal{G}_{1\mathbf{Y}_1^*}$ (the induced subgraph of $\mathcal{G}_1$) is intrinsic in $\mathcal{G}_1$, and every bidirected component in $\mathcal{G}_{t+1,t\mathbf{Y}_i^*}$ (the induced subgraph of $\mathcal{G}_{t+1,t}$) is intrinsic in $\mathcal{G}_{t+1,t}$, where $\mathbf{Y}_1^*$ is the set of ancestors of $\mathbf{Y} \cap \mathbf{V}_1$ not through $\mathbf{A} \cap \mathbf{V}_1$ in $\mathcal{G}_1$, and for every $i \in 2, \ldots, T$, $\mathbf{Y}_i^*$ is the set of ancestors of $\mathbf{Y} \cap \mathbf{V}_i$ not through $\mathbf{A} \cap \mathbf{V}_i$ in $\mathcal{G}_{t+1,t}$. Moreover, if $p(\mathbf{Y}(\mathbf{a}))$ is identified, we have*

$$\left(\sum_{\mathbf{Y}_1^* \setminus ((\mathbf{Y} \cup \mathbf{A}) \cap \mathbf{V}_1)} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1\mathbf{Y}_1^*})} q_{\mathbf{D}}^1(\mathbf{D} | \operatorname{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})|_{\mathbf{A}=\mathbf{a}} \right) \times$$

$$\prod_{i=2}^{T} \left(\sum_{\mathbf{Y}_i^* \setminus ((\mathbf{Y} \cup \mathbf{A}) \cap \mathbf{V}_i)} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{t+1,t\mathbf{Y}_i^*})} q_{\mathbf{D}}^{t+1,t}(\mathbf{D} | \operatorname{pa}_{\mathcal{G}}(\mathbf{D}) \setminus \mathbf{D})|_{\mathbf{A}=\mathbf{a}} \right),$$

where $q_{\mathbf{D}}^1$ and $q_{\mathbf{D}}^{t+1,t}$ are kernels corresponding to intrinsic sets that are districts in $\mathcal{D}(\mathcal{G}_{1\mathbf{Y}_1^*})$ and $\mathcal{D}(\mathcal{G}_{t+1,t\mathbf{Y}_1^*})$ in the nested Markov factorizations of $\mathcal{G}_1$ and $\mathcal{G}_{t+1,t}$, respectively.

This result, unlike in [Blondel et al., 2017], allows arbitrary sets of treatments in a DBN. The proof and an example are presented in Section B of the Appendix. If Assumption 1 does not hold, causal effects in causal DBNs may still be identified for any finite T (Lemma 1 in the Appendix), Section 4.2 in [Blondel et al., 2017]. However, the resulting functional will likely be computationally intractable.

4 FULLY OBSERVED PDSEMS

A crucial modeling assumption employed by causal DBNs is that both structure and parameterization remain invariant over time. This is ill-suited to capture the sort of path dependence described in the introduction. We now describe our approach to relaxing this assumption via path dependent structural equation models (PDSEMs), a generalization of causal DBNs capturing path dependence. A PDSEM may also be viewed as a generalization of a Markov decision process that explicitly represents causal relationships, including confounding, among variables that make up individual states in a process. See Section D.2 of the Appendix for details.

4.1 A SIMPLE PDSEM

To illustrate PDSEMs, we use a simple example inspired by the surgery setting. We assume a surgery will consist of three states: s^1 ("incision"), s^2 ("modification of bone/tissue"), and s^3 ("closing the incision"). Further, each state has the following variables: A (patient status prior to any procedures in the current stage), B (experience of surgeon performing the procedure in the current stage) and C (the observed patient outcome for the stage after procedure is performed), all observed. The surgery always starts at s^1, and concludes upon reaching s^3. Procedures performed in s^2 may either succeed, leading to s^3, or fail with some probability, leading the surgeon to revisit s^1. The state transition diagram for this scenario is shown in Fig. 4 (b).

The causal diagram in Fig. 4 (a) shows relationships between variables in s^1 and functions similar to the prior network in a causal DBN. In addition to variables A_1, B_1 and C_1, it contains S_1, representing the state to transition to at time step 1. In our simple model, the state s^1 transitions to s^2 with probability 1, and so S_1 represents a degenerate probability distribution and does not depend on any other variable. In general, however, the probability associated with S_1 may depend on other variables in the current state. Transitions are specified by multiple causal CDAGs, one for every allowed state transition. These CDAGs are shown in Fig. 4(c),(d) and (e) (where dashed edges are ignored). These graphs include transition edges representing relationships between variables in the state at time t and variables in the state at time $t + 1$, and state-specific relationships among variables at time $t + 1$ We assume state spaces of variables associated with each state are the same across state transition and prior graphs. For example, the state spaces of A_1, B_1, C_1 in Fig. 4(a) and A_{21}, B_{21}, C_{21} in Fig. 4(c) are the same, but the variables themselves (and the causal graphs relating them) are not. This implies values may be indexed by state, e.g. a_1 can refer without loss of generality to a value of A_1 or A_{21}. Similarly, conditional distributions that depend on variables in a prior state are well-defined if those variables are indexed by the prior state only, e.g. $p(A_{12}|A_1)$ is a shorthand for "a density over A_{12} in transition $(1, 2)$ given any value a_1 of any variable of the form A_{i1}." Causal graphs in 4(a),(c),(d),(e), along with the state-transition diagram 4(b), completely describe the fully observed PDSEM. Complex state dynamics are captured by distinct state causal DAGs and path-dependence is a consequence of state transitions that may depend on variables in the current state, and not

[2] The nested Markov factorization generalizes to CADMGs in the natural way, see Section B of the Appendix for details.

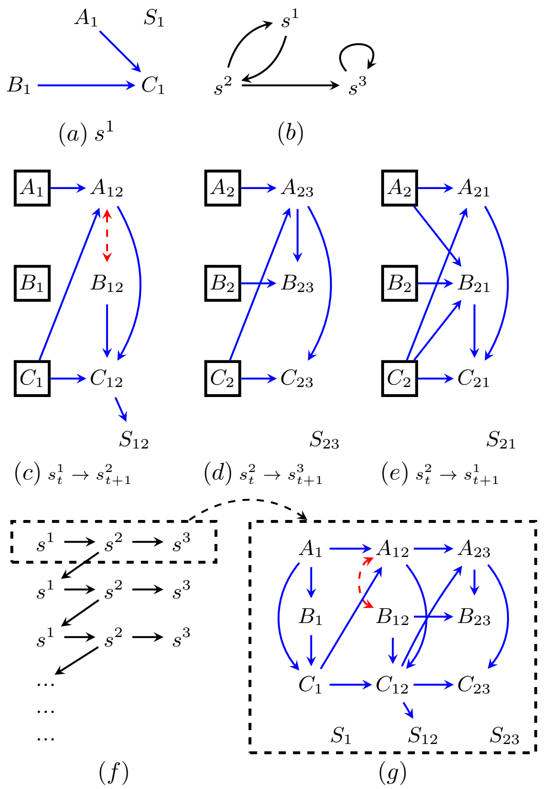

Figure 4: A simple PDSEM. (a) Causal structure of the initial state S^1. (b) The state transition diagram. (c),(d),(e) Causal diagrams representing possible transitions and subsequent states. (f) Causal relationships in a system evolving according to the state transitions: $s^1 \to s^2 \to s^3$. (g) A snapshot of a possible PDSEM trajectory that terminates in 3 timepoints is represented as an unrolled ADMG.

just the state itself.

The model we describe represents a randomized controlled trial where the surgeon operating during state s^2 is randomly assigned, hence B_{12} in the transition graph in Fig. 4 (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^3, B does not influence C, since closing the incision is a task adequately performed independent of surgeon experience. The observed data factorization of a fully-observed PDSEM is not finite, but yields a well defined joint distribution p_∞ over possible trajectories shown schematically in 4(f):

$$p_1 \prod_{t=1}^{\infty} (p_{12})^{\mathbb{I}(s_t^1, s_{t+1}^2)} \, (p_{23})^{\mathbb{I}(s_t^2, s_{t+1}^3)} \, (p_{21})^{\mathbb{I}(s_t^2, s_{t+1}^1)} \, 1^{\mathbb{I}(s_t^3)}$$
$$p_1 = p(A_1)p(B_1|A_1)p(C_1|A_1,B_1)\tilde{p}(S_1)$$
$$p_{12} = p(A_{12}|A_1,C_1)p(B_{12})p(C_{12}|B_{12},A_{12},C_1)p(S_{12}|C_{12})$$
$$p_{23} = p(A_{23}|A_2,C_2)p(B_{23}|B_2,A_{23})p(C_{23}|A_{23},C_2)\tilde{p}(S_{23})$$
$$p_{21} = p(A_{21}|A_2,C_2)p(B_{21}|B_2,A_2,C_2)p(C_{21}|C_2,B_{21},A_{21})\tilde{p}(S_{21}),$$

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.

PDSEMs allow us to reason about counterfactual questions such as: "what would happen if all procedures are performed by the resident surgeon ($B = b$), possibly contrary to fact?". The counterfactual joint distribution $p_\infty(b)$ is obtained by standard structural equation replacement semantics [Pearl, 2009], on the state-specific marginal and conditional counterfactual distributions:

$$p_1(b) \prod_{t=1}^{\infty} (p_{12}(b))^{\mathbb{I}(s_t^1, s_{t+1}^2)} (p_{23}(b))^{\mathbb{I}(s_t^2), s_{t+1}^3)} (p_{21}(b))^{\mathbb{I}(s_t^2, s_{t+1}^1)} 1^{\mathbb{I}(s_t^3)},$$

which is identified by using the g-formula for every component of the factorization, in a generalization of (4), yielding:

$$p_0^* \prod_{t=1}^{\infty} (p_{12}^*)^{\mathbb{I}(s_t^1, s_{t+1}^2)} \, (p_{23}^*)^{\mathbb{I}(s_t^2), s_{t+1}^3)} \, (p_{21}^*)^{\mathbb{I}(s_t^2, s_{t+1}^1)} \, 1^{\mathbb{I}(s_t^3)}$$
$$p_1^* = p(A_1)p(C_1|A_1, b)\tilde{p}(S_1)$$
$$p_{12}^* = p(A_{12}|A_1, C_1)p(C_{12}|b, A_{12}, C_1)p(S_{12}|C_{12})$$
$$p_{23}^* = p(A_{23}|A_2, C_2)p(C_{23}|A_{23}, C_2)\tilde{p}(S_{23})$$
$$p_{21} = p(A_{21}|A_2, C_2)p(C_{21}|C_2, b, A_{21})\tilde{p}(S_{21}).$$

While the distribution $p(S_{12}|C_{12})$ remains the same, the probability that s^1 is visited from s^2 is likely higher in $p_\infty(b)$ compared to p_∞. This is because B_{12}, counterfactually set to b, causes C_{12}, and C_{12} causes S_{12}. Thus, PDSEMs encode counterfactually changing state transition probabilities from their observed values.

4.2 AN ARBITRARY PDSEM

An arbitrary PDSEM is defined using a set of states $\mathbf{s}$, with initial state s^1, an absorbing state s^*, a set $\mathcal{T}$ of state index pairs of the form (i, j), where $s^i \neq s^*$ representing allowed state transitions, a DAG $\mathcal{G}_1$ on $\mathbf{V}_1$ for the initial state s^1, and for each $(i, j) \in \mathcal{T}$, a CDAG $\mathcal{G}_{ij}$ on $\mathbf{V}_{ij}$ given $\mathbf{V}_i$. Variables $S_1 \in \mathbf{V}_1, \{S_{ij} \in \mathbf{V}_{ij} : (i, j) \in \mathcal{T}\}$ determine probabilities of transitioning from state to state. Just as in a causal DBN, the DAG $\mathcal{G}_1$, and CDAGs $\mathcal{G}_{ij}$ represent structural equation models for the initial state, and the appropriate state transitions, respectively. That is, in the initial state, each variable $V \in \mathbf{V}_1$ is determined via $f_V(\text{pa}_{\mathcal{G}}(V), \epsilon_V)$. Similarly, for each variable $V \in \mathbf{V}_{ij}$ in any state transition represented by $\mathcal{G}_{ij}$. We assume $S_1, \{S_{ij} : (i, j) \in \mathcal{T}\}$ have no outgoing edges (this is without loss of generality, as structural equations are already state-specific in a PDSEM).

We note that while we define PDSEMs using the structural model semantics [Pearl, 2009], we only consider identification and estimation of interventional distributions, which only requires that interventional distributions in the full data causal model may be represented by a truncated factorization. Thus, the model we consider may be viewed as a path-dependent version of a causal Bayesian network [Pearl, 2009], or the causal model described in [Spirtes et al., 2001]. Considering path-dependence for causal inference problems that require the full generality of the structural equation formalism, such as mediation analysis, is an interesting extension for future work.

A first order Markov PDSEM obeys the following assumption that ensures that we need not condition on any context in the past except variables in the prior state.

Assumption 2 *For every state s^j, any CDAG $\mathcal{G}_{ij}$ or DAG $\mathcal{G}_j$ will have random variables that share state spaces.*

We thus denote the values of any $\mathbf{V}_{ij}$ for any transition (i, j) into state j by $\mathbf{v}_j$ (note the lack of dependence on i). As in our example, we index conditional densities that depend on variables in a prior state by that state only, e.g. $p(A_{12}|A_1)$.

Define $\mathbf{V} \equiv \mathbf{V}_1 \cup \left(\bigcup_{(i,j)\in\mathcal{T}} \mathbf{V}_{ij}\right)$. A PDSEM yields an observed distribution $p_\infty(\mathbf{V})$ with the factorization:

$$p_1(\mathbf{V}_1) \prod_{t=1}^{\infty} \left(\prod_{(i,j)\in\mathcal{T}} (p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i))^{\mathbb{I}(s_t^i, s_{t+1}^j)} \right) 1^{\mathbb{I}(s_t^*)} \tag{6}$$

where $p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i) = \prod_{V\in\mathbf{V}_{ij}} p(V|\,\mathrm{pa}_{\mathcal{G}_{ij}}(V))$ and $p_1(\mathbf{V}_1) = \prod_{V\in\mathbf{V}_1} p(V|\,\mathrm{pa}_{\mathcal{G}_1}(V))$.

An intervention in a PDSEM is defined on a set of treatment variables $\mathbf{A} \equiv \bigcup_{(i,j)\in\mathcal{T}} \mathbf{A}_{ij}$ and set to values $\mathbf{a}$ with the property that for any $(i, j), (k, j) \in \mathcal{T}$, the same values $\mathbf{a}_j$ are being set to $\mathbf{A}_{ij}$ and $\mathbf{A}_{ij}$. Define $\mathbf{Y}_{ij}$ in each transition graph $\mathcal{G}_{ij}$ to be all variables in that state not in $\mathbf{A}_{ij}$, with their corresponding values being $\mathbf{y}_j$, their union being $\mathbf{Y}$, and the values of the union being $\mathbf{y}$.

A new counterfactual distribution $p_\infty(\mathbf{Y}(\mathbf{a}))$ is obtained from the counterfactual initial state distribution $p_1(\mathbf{Y}_1(\mathbf{a}_1))$, and transition distributions $p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i))$ as:

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \prod_{t=1}^{\infty} \left(\prod_{(i,j)\in\mathcal{T}} (p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i)))^{\mathbb{I}(s_t^i, s_{t+1}^j)} \right) 1^{\mathbb{I}(s_t^*)}$$

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 DBN g-formula (4) to PDSEMs.

Lemma 2 *Given a fully observed PDSEM, each factor of the distribution $p_\infty(\mathbf{Y}(\mathbf{a}))$ is identified from $p_\infty(\mathbf{V})$ as:*

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \equiv \prod_{V\in\mathbf{Y}_1\setminus\mathbf{A}_1} p_1(V|\,\mathrm{pa}_{\mathcal{G}_1}(V))\Big|_{\mathbf{A}_1=\mathbf{a}_1}$$

$$p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i)) \equiv \prod_{V\in\mathbf{Y}_{ij}\setminus\mathbf{A}_j} p_{ij}(V|\,\mathrm{pa}_{\mathcal{G}_{ij}}(V))\Big|_{\substack{\mathbf{A}_i=\mathbf{a}_i,\\ \mathbf{A}_j=\mathbf{a}_j}} \tag{7}$$

Just as with DBNs, a PDSEM may be generalized from a first order to a kth-order Markov model, where variables in a particular state, can depend on variables in at most k prior states. This involves an appropriate generalization of Assumption 2, and specification of a larger set of transition networks. Details are in Section C.2 od the Appendix.

If all transition networks in a PDSEM obey a single consistent topological order, it is possible to encode a PDSEM by a causal DBN. Such an encoding will be inefficient and non-intuitive, however, since this causal DBN would represent restrictions of a PDSEM via context-specific independences in a large transition network representing a Cartesian product of possible transition networks of a PDSEM. If a consistent topological order on variables in transition networks does not exist, PDSEMs do not have a known causal DBN representation. Details are in Section D.1 of the Appendix.

5 PDSEM WITH HIDDEN VARIABLES

In extending causal inference to latent variable PDSEMs, in addition to Assumption 1 and Assumption 2, we assume the probabilities of any state transition trajectories are observed.

Assumption 3 *The variables S_{ij} for any $(i, j) \in \mathcal{T}$ governing state transition probabilities are observed.*

The latent variable PDSEMs then decompose into an initial state and a set of transitions such that causal inference results may be stated without loss of generality using latent projection ADMGs (and CADMGs) of appropriate DAGs and CDAGs. In addition, the fact that variables S_{ij} are observed implies we can evaluate counterfactual state transition probabilities, provided they are identified.

Formally, fix a PDSEM defined given the initial state DAG is $\mathcal{G}$ on $\mathbf{V}_1, \mathbf{H}_1$ and the set of transition CDAGs $\mathcal{G}_{ij}$ on $\mathbf{V}_{ij}, \mathbf{H}_{ij}$ given $\mathbf{V}_i$, for all $(i, j) \in \mathcal{T}$, such that: (i) the variables $\mathbf{V} \equiv \{\mathbf{V}_1\} \cup \bigcup_{(i,j)\in\mathcal{T}} \mathbf{V}_{ij}$, and $\mathbf{H} \equiv \{\mathbf{H}_1\} \cup \bigcup_{(i,j)\in\mathcal{T}} \mathbf{H}_{ij}$ are observed, and hidden, respectively, (ii) all state transition variables are observed ($S_1 \in \mathbf{V}_1$, $S_{ij} \in \mathbf{V}_{ij}$ for every $(i, j) \in \mathcal{T}$), and (iii) every state has the same observed and hidden variables regardless of transition (for every j and all $(i, j), (k, j) \in \mathcal{T}$, $\mathbf{H}_{ij} = \mathbf{H}_{kj}$ and $\mathbf{V}_{ij} = \mathbf{V}_{kj}$).

Given this definition of a latent variable PDSEM, the observed data distribution $p_\infty(\mathbf{V})$ is obtained from applying the usual transition probabilities to the margin at the initial state $p_1(\mathbf{V}_1) \equiv \sum_{\mathbf{H}_1} p_1(\mathbf{V}_1 \dot{\cup} \mathbf{H}_1)$, and the margins of all transition probabilities $p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i) \equiv \sum_{\mathbf{H}_{ij}} p_{ij}(\mathbf{V}_{ij} \dot{\cup} \mathbf{H}_{ij}|\mathbf{V}_i)$.

Fix a set of observed treatment variables $\mathbf{A}$, the union of $\{\mathbf{A}_{ij} : (i,j) \in \mathcal{T}\}$, such that $\mathbf{a}_j$ are set to $\mathbf{A}_{ij}, \mathbf{A}_{kj}$ for any $(i,j), (k,j) \in \mathcal{T}$, and the set of outcomes $\mathbf{Y}_{ij} = \mathbf{V}_{ij} \setminus \mathbf{A}_{ij}$ for any $(i,j) \in \mathcal{T}$, with $\mathbf{Y}$ the union of $\{\mathbf{Y}_{ij} : (i,j) \in \mathcal{T}\}$.

Identification for $p_\infty(\mathbf{Y}(\mathbf{a}))$ in a latent variable PDSEM reduces to identification theory for $p_1(\mathbf{Y}_1(\mathbf{a}_1))$ in the latent projection ADMG $\mathcal{G}_1$ on $\mathbf{V}_1$, and $p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{V}_i(\mathbf{a}_i))$ in the latent projection CADMG $\mathcal{G}_{ij}$ on $\mathbf{V}_{ij}$ given $\mathbf{V}_i$, as follows:

Lemma 3 *Under Assumptions 1, 2 and 3, given a latent variable PDSEM represented by $\mathcal{G}_1$ and $\{\mathcal{G}_{ij} : (i,j) \in \mathcal{T}\}$, $p_\infty(\mathbf{Y}(\mathbf{a}))$ is identified from $p_\infty(\mathbf{V})$ if and only if every bidirected component in $\mathcal{G}_{1\mathbf{Y}_1}$ is intrinsic in $\mathcal{G}_1$, and every bidirected component in $\mathcal{G}_{ij\mathbf{Y}_j}$ is intrinsic in $\mathcal{G}_{ij}$ for every i and j. Moreover, if $p_\infty(\mathbf{Y}(\mathbf{a}))$ is identified, it is equal to*

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) \prod_{t=1}^{\infty} \left(\prod_{(i,j) \in \mathcal{T}} (p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i)))^{\mathbb{I}(s^i_{t-1}, s^j_t)} \right) 1^{\mathbb{I}(s^*_{t-1})} \tag{8}$$

where

$$p_1(\mathbf{Y}_1(\mathbf{a}_1)) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1\mathbf{Y}^*_1})} q^1_{\mathbf{D}}(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_1}(\mathbf{D}))\Big|_{\mathbf{A}_1 = \mathbf{a}_1}, \tag{9}$$

where each kernel $q^1_{\mathbf{D}}(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_1}(\mathbf{D}))$ is in the nested Markov factorization of $p_1(\mathbf{V}_1)$ with respect to $\mathcal{G}_1$, and

$$p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a}_i)) = \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{V}_{ij} \setminus \mathbf{A}_{ij}})} q^{ij}_{\mathbf{D}}(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{ij}}(\mathbf{D}))\Big|_{\substack{\mathbf{A}_i = \mathbf{a}_i, \\ \mathbf{A}_j = \mathbf{a}_j}} \tag{10}$$

where each kernel $q^{ij}_{\mathbf{D}}(\mathbf{D}|\operatorname{pa}^s_{\mathcal{G}_{ij}}(\mathbf{D}))$ is in the nested Markov factorization of $p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i)$ with respect to $\mathcal{G}_{ij}$.

An example of a hidden variable PDSEM and identifying functionals are given in Section B of the Appendix. If Assumption 1 is violated, path dependence makes identification complicated in PDSEMs (Section C.3 of the Appendix).

6 EXPERIMENTS

6.1 SIMULATION STUDY

We simulate data and perform statistical inference using the PDSEM shown in Fig. 4 in Section 4.1. Details on inference are in Section F.1 of the Appendix. The system has states $\{s^1, s^2, s^3\}$ and variables $\{A, B, C\}$ in each state. Additionally, s^2 has a hidden common cause of A and B. This is represented by the red (dotted) bidirected edge $A \leftrightarrow B$ in the latent projected ADMG in Fig. 4(c). Patient health status A, surgeon experience B, and duration of the stage of surgery C, are all continuous variables. State and transition graphs are identical to those in Fig. 4. This PDSEM was used to consider the causal impact of surgeon experience (measured by total operating time in their career) on average surgery length. This outcome is easy to measure, and is known to serve as an informative proxy for other measures of surgery quality, such as follow-up assessments of quality of life [Rambachan et al., 2013, Jackson et al., 2011].

Parameters associated with the given generative model are $p(S_{t+1} = s^j|S_t = s^i, \mathbf{V}_t)$, where $s^i_t \to s^j_{t+1}$ is a transition allowed by the model, and $p(V^{ij}_{t+1} = v|S_{t+1} = s^j, S_t = s^i, \mathbf{V}_t)$, where $V^{ij} \in \{A^{ij}, B^{ij}, C^{ij}\}$, where $s^i_t \to s^j_{t+1}$ is an allowed transition. These are chosen to be reasonable for the surgery application, yielding a distribution Markov relative to appropriate graphs. We simulated $N = 10000$ "surgeries," with initial state s^1. Transition probabilities were generated using a logistic regression on variables in the current state, with transitions eventually terminating at the absorbing state. Each variable V_i is generated from a set of linear structural equations with correlated errors. Using generated data, state transition probabilities were estimated using maximum likelihood. Parameters for the structural equation model were estimated using the RICF algorithm [Drton et al., 2009], implemented in the Ananke package [Bhattacharya et al.].

We assessed the causal impact of surgeon experience on operating time by generating two sets of sampled surgery trajectories where, in each stage of the surgery, the surgeon was intervened to have higher (vs. lower) career operating time by one unit. These trajectories may be viewed as a Monte Carlo sampling scheme for evaluating the functional given by (8), (9) and (10). This approach generalizes similar schemes developed for longitudinal causal models [Westreich et al., 2012]. The comparison of these two sets of trajectories may be viewed as a generalization of the *average causal effect (ACE)* from classical longitudinal causal models to PDSEMs.

The results are shown in Fig. 5. Surgeries performed by experienced surgeons are shorter ($\mu = 5.79$, $\mathbf{q}_{0.05} = 3, \mathbf{q}_{0.95} = 13$) than those performed by trainees ($\mu = 7.02$, $\mathbf{q}_{0.05} = 3$, $\mathbf{q}_{0.95} = 17$) where $\mathbf{q}_p$ denotes the pth quantile. Surgeries performed by trainees have higher variance.

6.2 DATA APPLICATION

We are interested in the causal impact of surgeon experience on the average length of surgery, in the context of septoplasty. Such surgeries are characterized by multiple phases involving different tools and procedures, and the causal dynamics between variables differs from phase to phase. Moreover, surgeries do not always proceed sequentially, and may return to earlier phases depending on what happened in a particular phase (see Appendix F). To accurately model surgery lengths, we thus need to allow for a

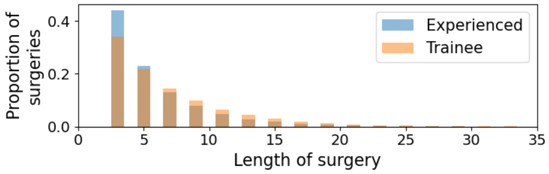

Figure 5: Histograms of the number of transitions in a surgery under two different interventions: when a more experienced surgeon performs the entire procedure, and when a less experienced trainee performs the entire procedure.

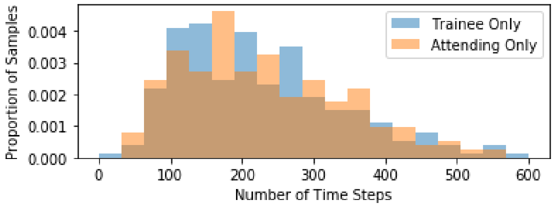

Figure 6: Histograms of hypothetical surgeries performed only by a junior trainee surgeon (blue) versus hypothetical surgeries performed only by a senior attending surgeon (orange). Surgeries performed by the attending are slightly longer ($\mu = 244.3.91, \sigma = 139.9$) than those of the trainee ($\mu = 233.5, \sigma = 125.9$).

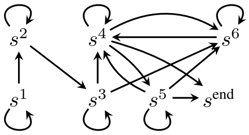

Figure 7: The state transition diagram for the surgery data application.

variety of phases, and complex structure within each phase, making PDSEM a suitable model.

Our dataset consists of 236 septoplasty procedures conducted at our institution's research hospital. A total of 57343 timestamped records were collected, including 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). An artificial absorbing state s^{end} represents the end of procedures. 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 by a team. Also, attending surgeons perform for 64.98% of all operating time and trainees the rest. Twelve different surgical tools were tracked for use. The state transition diagram representing allowed state transitions is presented in Fig 7. We discretized all variables into two categories, and fit model parameters by maximum likelihood.

While there are certainly unobserved but relevant confounding variables in the problem we consider (such as underlying patient state), we assume these variables influence treatment variables (identity of the surgeon), as well as variables in the next stage only via relevant observed variables (such as duration of the stage, and tools currently in use). In addition to implying Assumption 1, this implies identifiability of the parameter of interest (a contrast of the average length of surgery had experienced vs inexperienced surgeon performed all stages) is given by Lemma 2, and statistical inference may be performed as if the prior network were a DAG, and every transition network were a CDAG, without loss of generality. An interesting area of future work is generalizing sensitivity analysis methods developed in classical causal models for assessing robustness to violations of the lack of unobserved confounding assumptions to PDSEMs. PDSEMs that arise when Assumption 1 is relaxed are discussed in Section C.2 of the Appendix.

Estimation of $p(s_t|s_{t-1}, \mathbf{v}_{t-1})$ at all levels of $s_{t-1}, \mathbf{v}_{t-1}$ is not always possible due to finite sample limitations. To address this, we apply additive smoothing to $p(s_t|s_{t-1}, \mathbf{v}_{t-1})$, based on the empirical distribution $p(s_t|s_{t-1})$. Goodness of fit is illustrated in Fig 4 of the Appendix and results are presented in Fig 6. 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 our learned parameters, is close to zero. This indicates that policies that govern the trade-off between the need to train surgeons, and overall surgery quality (as quantified by our outcome) are effective at our institution.

7 CONCLUSIONS

In this paper, we have introduced the Path Dependent Structural Equation Model (PDSEM) for longitudinal data unifying complex state structure from DBNs and complex state transition dynamics from MDPs. It can also be seen as a graphical model generalizing 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 estimation. We showed the utility of the model in clinical settings using simulations as well as data from a septoplasty 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.

Acknowledgements

This project was supported by ONR N00014-18-1-2760, NSF CAREER 1942239, NSF EAGER 1939675, and R01 AI127271-01A1. The data used in this paper were collected as part of research supported by grants NIH; R21-

DE022656 and R01-DE05265 (PI: Dr. Masaru Ishii, MD, PhD). We thank Dr. Narges Ahmidi from the Institute of Computational Biology, Helmholtz Munich Center and Dr. Greg Hager, Dr. Masaru Ishii, Dr. Swaroop Vedula and the rest of their team at Johns Hopkins University for access to this data and their valuable insights on surgical training.

References

Narges Ahmidi, Piyush Poddar, Jonathan D. Jones, Swaroop S. 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.

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

Rohit Bhattacharya, Jaron J. R. Lee, Razieh Nabi, and Ilya Shpitser. *Ananke*: A python package for causal inference with graphical models. URL `https://ananke.readthedocs.io/en/latest/index.html`.

Gilles Blondel, Marta Arias, and Ricard Gavaldà. Identifiability and transportability in dynamic causal networks. *International Journal of Data Dcience and Analytics*, 3 (2):131–147, 2017.

Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(10), 2009.

Miguel Hernan and James M. Robins. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.

Miguel A. 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.

Timothy D. Jackson, Jeffrey J. Wannares, Todd R. 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.

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

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 nonstationary vector autoregression. *Proceedings of Machine Learning Research*, 89:2986–2994, April 2019. ISSN 2640-3498.

Søren W. Mogensen, Daniel Malinsky, and Niels R. Hansen. Causal learning for partially observed stochastic dynamical systems. In *UAI*, pages 350–360, 2018.

Kevin P. Murphy. *Machine learning: A probabilistic perspective*. MIT Press, September 2012. ISBN 978-0-262-30432-0.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–709, 1995. URL `citeseer.ist.psu.edu/55450.html`.

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. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

Mark A. Richards. *Fundamentals of radar signal processing*. Tata McGraw-Hill Education, 2005.

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

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer Verlag, New York, 2 edition, 2001. ISBN 978-0262194402.

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

Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.

Daniel Westreich, Stephen R. Cole, Jessica G. Young, Frank Palella, Phyllis C. Tien, Lawrence Kingsley, Stephen J. Gange, and Miguel A. Hernán. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident aids or death. *Statistics in Medicine*, 31 (18):2000–2009, 2012.

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical Report R-23, Purdue AI Lab, 2016.