# Path Dependent Structural Equation Models - Supplementary Material

**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

## A GRAPH PRELIMINARIES

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

Standard genealogic relations on graphs are as follows: parents, children, descendants, siblings and ancestors of $X$ in a graph $\mathcal{G}$ are denoted by $\text{pa}_\mathcal{G}(X), \text{ch}_\mathcal{G}(X), \text{de}_\mathcal{G}(X), \text{si}_\mathcal{G}(X), \text{an}_\mathcal{G}(X)$, respectively [Lauritzen, 1996]. These relations are defined disjunctively for sets, e.g. $\text{pa}_\mathcal{G}(\mathbf{X}) \equiv \bigcup_{X \in \mathbf{X}} \text{pa}_\mathcal{G}(X)$. By convention, for any $X$, $\text{an}_\mathcal{G}(X) \cap \text{de}_\mathcal{G}(X) \cap \text{dis}_\mathcal{G}(X) = \{X\}$.

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

In the subsequent discussion, we will denote an ADMG $\mathcal{G}$ on $\mathbf{V}$ by notation $\mathcal{G}(\mathbf{V})$, and a CADMG $\mathcal{G}$ on $\mathbf{V}$ given $\mathbf{W}$ by notation $\mathcal{G}(\mathbf{V}, \mathbf{W})$.

## B THE NESTED MARKOV FACTORIZATION

### B.1 WHY DO WE NEED AN ALTERNATIVE FACTORIZATION?

A hidden variable CDAG $\mathcal{G}(\mathbf{V} \cup \mathbf{H}, \mathbf{W})$ may be used to define a factorization on distributions $p(\mathbf{V}|\mathbf{W})$ in terms of the CDAG as: $p(\mathbf{V}|\mathbf{W}) = \sum_\mathbf{H} \prod_{V \in \mathbf{V} \cup \mathbf{H}} p(V | \text{pa}_\mathcal{G}(V))$. 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, 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 $p(\mathbf{V}|\mathbf{W})$ can be defined directly on the latent projection CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$. 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 $p(\mathbf{V}|\mathbf{W})$ [Shpitser et al., 2018]. In addition, $p(Y(a)|\mathbf{W})$ (an interventional distribution given a fixed context $\mathbf{W}$) identified in a hidden variable causal model represented by $\mathcal{G}(\mathbf{V} \cup \mathbf{H}, \mathbf{W})$ is always equal to a modified version of a nested factorization [Richardson et al., 2017] associated with $\mathcal{G}(\mathbf{V}, \mathbf{W})$, described here.

### B.2 THE NESTED MARKOV FACTORIZATION

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

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

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

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

$$q_{\mathbf{V}\setminus\{V\}}(\mathbf{V} \setminus \{V\}|\mathbf{W} \cup \{V\}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(V|\,\text{mb}_{\mathcal{G}}(V))}$$

**Fixing sequences:** A sequence $\langle V_1, \ldots, V_k\rangle$ is said to be *valid* in $\mathcal{G}(\mathbf{V},\mathbf{W})$ if $V_1$ fixable in $\mathcal{G}(\mathbf{V},\mathbf{W})$, $V_2$ is fixable in $\phi_{V_1}(\mathcal{G}(\mathbf{V},\mathbf{W}))$, and so on. If any two sequences $\sigma_1, \sigma_2$ for the same set $\mathbf{S} \subseteq \mathbf{V}$ are fixable in $\mathcal{G}$, they lead to the same CADMG. The graph fixing operator can be extended to a set $\mathbf{S}$: $\phi_{\mathbf{S}}(\mathcal{G})$. This operator is defined as applying the vertex fixing operation in any valid sequence $\sigma$ for set $\mathbf{S}$.

Given a sequence $\sigma_{\mathbf{S}}$, define $\eta(\sigma_{\mathbf{S}})$ to be the first element in $\sigma_{\mathbf{S}}$, and $\tau(\sigma_{\mathbf{S}})$ to be the subsequence of $\sigma_{\mathbf{S}}$ containing all elements but the first. Given a sequence $\sigma_{\mathbf{S}}$ on elements in $\mathbf{S}$ valid in $\mathcal{G}(\mathbf{V},\mathbf{W})$, the kernel fixing operator $\phi_{\sigma_{\mathbf{S}}}(q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}), \mathcal{G}(\mathbf{V},\mathbf{W}))$ is defined to be equal to $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ if $\sigma_{\mathbf{S}}$ is the empty sequence, and $\phi_{\tau(\sigma_{\mathbf{S}})}(\phi_{\eta(\sigma_{\mathbf{S}})}(q_{\mathbf{V}}(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V},\mathbf{W})), \phi_{\eta(\sigma_{\mathbf{S}})}(\mathcal{G}(\mathbf{V},\mathbf{W})))$ otherwise.

**Reachability:** Given a CADMG $\mathcal{G}(\mathbf{V},\mathbf{W})$, a set $\mathbf{R} \subseteq \mathbf{V}$ is called *reachable* if there exists a sequence for $\mathbf{V} \setminus \mathbf{R}$ valid in $\mathcal{G}(\mathbf{V},\mathbf{W})$. In other words, if $\mathbf{S}$ is fixable in $\mathcal{G}$, $\mathbf{V} \setminus \mathbf{S}$ is reachable.

**Intrinsic sets:** A set $\mathbf{R}$ reachable in $\mathcal{G}(\mathbf{V},\mathbf{W})$ is *intrinsic* in $\mathcal{G}(\mathbf{V},\mathbf{W})$ if $\phi_{\mathbf{V}\setminus\mathbf{R}}(\mathcal{G})$ contains a single district, $\mathbf{R}$ itself. The set of intrinsic sets in a CADMG $\mathcal{G}$ is denoted by $\mathcal{I}(\mathcal{G})$.

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

$$\phi_{\sigma_{\mathbf{R}}}(p(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V},\mathbf{W})) = \prod_{\mathbf{D}\in\mathcal{D}(\phi_{\mathbf{R}}(\mathcal{G}(\mathbf{V},\mathbf{W})))} q_{\mathbf{D}}(\mathbf{D}|\,\text{pa}^{\text{s}}{}_{\mathcal{G}}(\mathbf{D}))$$

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

The *district factorization* or *Tian factorization* of $p(\mathbf{V}|\mathbf{W})$ results from the nested factorization:

$$\begin{aligned} p(\mathbf{V}|\mathbf{W}) &= \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}(\mathbf{V},\mathbf{W}))} q_{\mathbf{D}}(\mathbf{D}|\,\text{pa}^{\text{s}}{}_{\mathcal{G}}(\mathbf{D})) \\ &= \prod_{\mathbf{D}\in\mathcal{D}(\mathcal{G}(\mathbf{V},\mathbf{W}))} \left( \prod_{D\in\mathbf{D}} p(D \mid \text{pre}_{\prec}(D)) \right), \end{aligned}$$

where $\text{pre}_{\prec}(D)$ is the set of predecessors of $D$ according to a topological total ordering $\prec$. Each factor $\prod_{D\in\mathbf{D}} p(D \mid \text{pre}_{\prec}(D))$ is only a function of $\mathbf{D} \cup \text{pa}_{\mathcal{G}}(\mathbf{D})$ under the nested factorization.

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

### B.3 IDENTIFICATION

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

$$\begin{aligned} & p(\mathbf{Y}(\mathbf{a})|\mathbf{W}) \\ & = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} p(\mathbf{D}|\operatorname{do}(\mathrm{pa^s}_{\mathcal{G}}(\mathbf{D})))|_{\mathbf{A}=\mathbf{a}} \\ & = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{\mathbf{Y}^*})} \phi_{\mathbf{V} \setminus \mathbf{D}}(p(\mathbf{V}|\mathbf{W}); \mathcal{G}(\mathbf{V}, \mathbf{W}))|_{\mathbf{A}=\mathbf{a}} \end{aligned}$$

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

### B.4 EXAMPLE OF THE NESTED FACTORIZATION OF A HIDDEN VARIABLE PDSEM

A hidden variable PDSEM can be unrolled into a latent-projected ADMG if the model obeys restrictions given in Section 5. For instance, Fig. 1 in this Appendix shows an example where the first two states of the system involve hidden variables. In particular, the system at $s^2$ is the *front-door-graph* previously encountered in Section 2. Transition graphs are in Fig. 1(c)-(e).

The nested factorization for the initial graph in Fig. 1 (a) has intrinsic sets

$$(a) : \{A_1\}, \{B_1\}, \{C_1\}, \{A_1, B_1\}, \{S_1\}$$

with corresponding kernels

$$\begin{aligned} (a) : & \, q_{A_1}(A_1) \equiv p(A_1); q_{B_1}(B_1) = p(B_1); \\ & q_{C_1}(C_1|A_1, B_1) \equiv p(C_1|A_1, B_1); \\ & q_{A_1,B_1}(A_1, B_1) \equiv p(A_1, B_1); q_{S_1}(S_1) \equiv p(S_1). \end{aligned} \tag{1}$$

Similarly, the nested factorizations for the transition graphs in Fig. 1 (c),(d),(e) have intrinsic sets:

$$\begin{aligned} (c) &: \{A_{12}\}, \{B_{12}\}, \{C_{12}\}, \{A_{12}, C_{12}\}, \{S_{12}\} \\ (d) &: \{A_{23}\}, \{B_{23}\}, \{C_{23}\}, \{S_{23}\} \\ (e) &: \{A_{21}\}, \{B_{21}\}, \{A_{21}, B_{21}\}, \{C_{21}\}, \{S_{21}\}, \end{aligned}$$

with corresponding kernels

$$\begin{aligned} (c) : & \, q_{A_{12}}(A_{12}|C_1) \equiv p(A_{12}|C_1); q_{B_{12}}(B_{12}|A_{12})) \equiv p(B_{12}|A_{12}); \\ & q_{C_{12}}(C_{12}|C_1, B_{12}) \equiv \sum_{A_{12}} p(C_{12}|B_{12}, A_{12}, C_1) p(A_{12}|C_1); \\ & q_{A_{12},C_{12}}(A_{12}, C_{12}|B_{12}, C_1) \equiv p(C_{12}|B_{12}, A_{12}, C_1) p(A_{12}|C_1); q_{S_{12}}(S_{12}|C_{12}) \equiv p(S_{12}|C_{12}). \\ (d) : & \, q_{A_{23}}(A_{23}|A_2, C_2) \equiv p(A_{23}|A_2, C_2); q_{B_{23}}(B_{23}|B_2, A_{23}) \equiv p(B_{23}|B_2, A_{23}); \\ & q_{C_{23}}(C_{23}|C_2, A_{23}) \equiv p(C_{23}|C_2, A_{23}); q_{S_{23}}(S_{23}) \equiv p(S_{23}). \\ (e) : & \, q_{A_{21}}(A_{21}|A_2) \equiv p(A_{21}|A_2); q_{B_{21}}(B_{21}|B_2) \equiv p(B_{21}|B_2); q_{A_{21},B_{21}}(A_{21}, B_{21}) \equiv p(A_{21}, B_{21}); \\ & q_{C_{21}}(C_{21}|C_2, B_{21}, A_{21}) \equiv p(C_{21}|C_2, B_{21}, A_{21}); q_{S_{21}}(S_{21}) \equiv p(S_{21}). \end{aligned} \tag{2}$$

Applying the Nested Markov factorization on the trajectory in 1 (f), we obtain the following factorization:

$$\begin{aligned} & p(A_1, B_1, C_1) \cdot p(A_{12}, B_{12}, C_{12}|A_1, B_1, C_1) \cdot p(A_{23}, B_{23}, C_{23}|A_{12}, B_{12}, C_{12}) \\ & = \underbrace{\{q_{A_1,B_1}(A_1, B_1) q_{C_1}(C_1 \mid A_1, B_1)\}}_{(a)} \cdot \underbrace{\{q_{A_{12},C_{12}}(A_{12}, C_{12}|B_{12}, A_1, C_1) q_{B_{12}}(B_{12}|A_{12})\}}_{(c)} \cdot \\ & \underbrace{\{q_{A_{23}}(A_{23}|A_{12}, C_{12}) \cdot q_{A_{23}}(B_{23}|B_{12}, A_{23}) \cdot q_{C_{23}}(C_{23}|C_{12}, A_{23})\}}_{(d)}, \end{aligned}$$

where the kernels are given in (1) and (2) above.

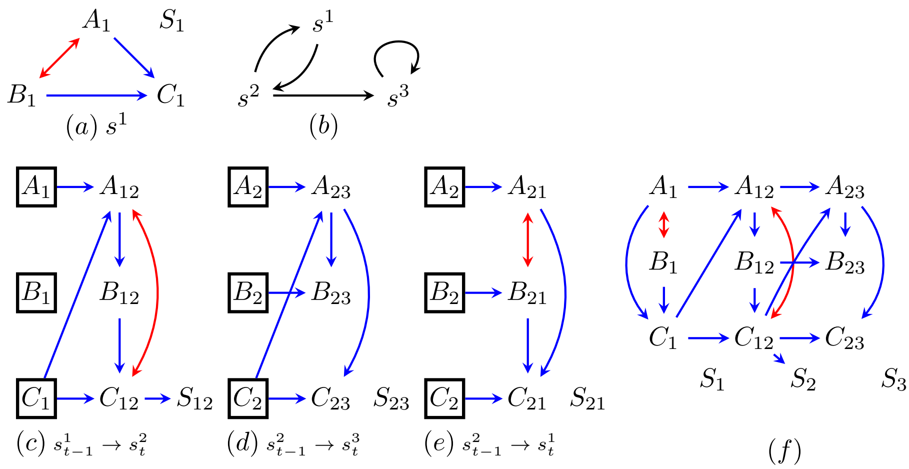


Figure 1: A hidden variable PDSEM. (a) Causal structure of the initial state $S^1$. (b) The state transition diagram. (c),(d),(e) Latent projected causal diagrams representing possible transitions and subsequent states. (f) A snapshot of a possible PDSEM trajectory represented as an unrolled ADMG

# C GENERALIZATIONS OF PDSEMS

## C.1 PDSEMS WITH OBSERVED CONTEXT

Here we describe how PDSEMs may be generalized to a setting where variables in the prior network and all transition networks depend on a vector of baseline covariates $\mathbf{W}$. We start by describing how an ordinary graphical causal model may be generalized in this way.

In a causal model $\mathcal{G}(\mathbf{V},\mathbf{W})$ with $Y \in \mathbf{V}$, $\mathbf{A} \subseteq \mathbf{V} \setminus \{Y\}$ and observed context (confounding) $\mathbf{W}$, counterfactuals are denoted by $Y(\mathbf{a})|\mathbf{W}$. Given a set $\mathbf{Y}$, $\mathbf{Y}(\mathbf{a})|\mathbf{W} \equiv \{Y(\mathbf{a})|\mathbf{W} : Y \in \mathbf{Y}\}$.

In an arbitrary PDSEM with observed variables $\mathbf{V}$ , and observed context $\mathbf{W}$, the pieces that define the model are a CDAG $\mathcal{G}_1(\mathbf{V}_1,\mathbf{W})$ for the initial state $s^1$, and for each transition $(i,j) \in \mathcal{T}$, a CDAG $\mathcal{G}_{ij}(\mathbf{V}_{ij}, \vec{V}_i \cup \mathbf{W})$

As before, we will impose Assumptions 1 and 2. Define $\mathbf{V} \equiv \mathbf{V}_1 \cup \left(\bigcup_{(i,j)\in\mathcal{T}} \mathbf{V}_{ij}\right)$. A PDSEM yields an observed data distribution $p_\infty(\mathbf{V},\mathbf{W})$ with the following factorization:

$$p(\mathbf{W})p_1(\mathbf{V}_1|\mathbf{W})\prod_{t=1}^{\infty}\left(\prod_{(i,j)\in\mathcal{T}}(p_{ij}(\mathbf{V}_{ij}|\mathbf{W}))^{\mathbb{I}(s^i_{t-1},s^j_t)}\right)1^{\mathbb{I}(s^*_{t-1})}$$

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

Since $\mathbf{W}$ is a set of context variables, we assume this set cannot be intervened on. We can define interventions $\mathbf{A} \equiv \bigcup_{(i,j)\in\mathcal{T}} \mathbf{A}_{ij}$, 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 $\mathbf{A}_{kj}$, the same as before. Define $\mathbf{Y}_{ij}$ in each transition graph $\mathcal{G}_{ij}$ to be all variables in that state not in $\mathbf{A}_{ij}$ or $\mathbf{W}$, with their corresponding values being $\mathbf{y}_j$, their union being $\mathbf{Y}$, and the values of the union being $\mathbf{y}$. This gives the new counterfactual distribution to be:

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

Thus, Lemma 2 in the main paper can be rewritten as follows:

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

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

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

where $p(\mathbf{W})$ is given as a part of the observed data distribution.

A PDSEM with hidden variables and observed context can be defined given the initial state CDAG is $\mathcal{G}$ on $\mathbf{V}_1, \mathbf{H}_1, \mathbf{W}$ and the set of transition CDAGs $\mathcal{G}_{ij}$ on $\mathbf{V}_{ij}, \mathbf{H}_{ij}, \mathbf{W}$ given $\mathbf{V}_i$, for all $(i,j) \in \mathcal{T}$, such that $S_1 \in \mathbf{V}_1$, $S_{ij} \in \mathbf{V}_{ij}$ for every $(i,j) \in \mathcal{T}$, for every $j$ and all $(i,j),(k,j) \in \mathcal{T}$, $\mathbf{H}_{ij} = \mathbf{H}_{kj}$ and $\mathbf{V}_{ij} = \mathbf{V}_{kj}$. We assume the variables $\mathbf{V} \equiv \{\mathbf{V}_1\} \cup \bigcup_{(i,j)\in\mathcal{T}} \mathbf{V}_{ij}$, and $\mathbf{W}$ are observed and $\mathbf{H} \equiv \{\mathbf{H}_1\} \cup \bigcup_{(i,j)\in\mathcal{T}} \mathbf{H}_{ij}$ are hidden.

The observed data distribution $p_\infty(\mathbf{V}, \mathbf{W})$ is obtained from $p(\mathbf{W})$, $p_1(\mathbf{V}_1|\mathbf{W}) \equiv \sum_{\mathbf{H}_1} p_1(\mathbf{V}_1 \dot\cup \mathbf{H}_1|\mathbf{W})$, and $p_{ij}(\mathbf{V}_{ij}|\mathbf{V}_i, \mathbf{W}) \equiv \sum_{\mathbf{H}_{ij}} p_{ij}(\mathbf{V}_{ij} \dot\cup \mathbf{H}_{ij}|\mathbf{V}_i, \mathbf{W})$. 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 $\{\mathbf{Y}_{ij} : (i,j) \in \mathcal{T}\}$. As before, we do not intervene on $\mathbf{W}$. Identification for $p_\infty(\mathbf{Y}(\mathbf{a}))$ in a latent variable PDSEM reduces to identification theory for $p_1(\mathbf{Y}_1(\mathbf{a}_1)|\mathbf{W})$ in the latent projection ADMG $\mathcal{G}_1(\mathbf{V}_1, \mathbf{W})$, and $p_{ij}(\mathbf{Y}_{ij}(\mathbf{a}_j)|\mathbf{V}_i(\mathbf{a}_i), \mathbf{W})$ in the latent projection CADMG $\mathcal{G}_{ij}$ on $\mathbf{V}_{ij}$ given $\mathbf{V}_i, \mathbf{W}$.

Lemma 3 in the main paper can be restated with observed context as follows:

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 $p_\infty(\mathbf{V}, \mathbf{W})$ 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 may be obtained from $p(\mathbf{W})$, which is a part of the observed data distribution, as well as:

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

where

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

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

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

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 $\mathcal{G}_{ij}$.

### C.2 $k$TH-ORDER MARKOV TEMPORAL CAUSAL MODELS

Both causal DBNs and PDSEMs may be generalized to $k$th-order Markov models, where variables in a particular time step depend on variables in at most $k$ prior states.

A $k$th-order Markov DBN consists of a single *prior network* $\mathcal{G}_1$, which is a DAG with vertices $\mathbf{V}_1$, a set of $k-1$ initial transition networks $\mathcal{G}_2, \ldots, \mathcal{G}_k$, where each $\mathcal{G}_i$ is a CDAG with random vertices $\mathbf{V}_i$ and fixed vertices $\bigcup_{j=1}^{i-1} \vec{V}_j$, and a transition network $\mathcal{G}_t$ with random vertices $\mathbf{V}_t$ and fixed vertices $\bigcup_{j=t-k}^{t-1} \vec{V}_j$. Each DAG and CDAG in a $k$th-order DBN is associated with a factorization of the corresponding joint or conditional distribution. The "unrolled" factorization of the DBN makes use of the prior distribution $p_1(\mathbf{V}_1)$ and initial transition network distributions $p_i(\mathbf{V}_i|\mathbf{V}_1, \ldots, \mathbf{V}_{i-1})$ for the first $k-1$ steps, and then uses a repeated version of the transition network distribution $p_t(\mathbf{V}_t|\mathbf{V}_{t-k}, \ldots, \mathbf{V}_{t-1})$:

$$\prod_{V\in\mathbf{V}_1} p_1(V|\operatorname{pa}_{\mathcal{G}_1}(V)) \prod_{\tilde{t}=1}^{k-1} \prod_{V\in\mathbf{V}_{\tilde{t}}} p_{\tilde{t}}(V|\operatorname{pa}_{\mathcal{G}_{\tilde{t}}}(V)) \cdot \prod_{\tilde{t}=k}^{T-1} \prod_{V\in\mathbf{V}_t} p_t(V|\operatorname{pa}_{\mathcal{G}_t}(V)). \tag{3}$$

The causal version of a $k$th-order Markov DBN is obtained in the natural way by endowing each DAG and CDAG with structural equation model semantics, and obtaining standard identification results, via the g-formula, and the ID algorithm in cases hidden variables are present.

The relaxation of the first-order Markov assumption in these models does not come without a cost: additional transition networks must be specified, and all transition networks may potentially depend on a larger set of variables, resulting in a more difficult statistical inference problem on model parameters.

PDSEMs may similarly be relaxed to a $k$th-order Markov model. For example, given a model with 3 states, if we wish all transitions to depend on two rather than one prior state, we would need to specify a prior network (with a corresponding causal model), a set of 3 single-step transition networks (corresponding to steps from the initial state to any of the 3 possible states), and then finally a set of 9 transition networks, representing variables in one of three states that depend on any two prior states (which may involve states repeating). Such a model would have a separate transition network $\mathcal{G}_{\langle 1,2,3\rangle}$ for variables in state 3 at time $t$, where state 2 was visited at time $t-1$, and state 1 was visited at time $t-2$, and a transition network $\mathcal{G}_{\langle 2,1,3\rangle}$ for variables in state 3 at time $t$, where state 1 was visited at time $t-1$, and state 2 was visited at time $t-2$.

In general, a $k$th-order Markov PDSEM with $S$ states will have a single prior network DAG corresponding to the initial state, $S^i$ transition networks CDAGs that depend on $i$ prior states (for $i = 1, \ldots, k-1$), indexed by sequences of states visited (starting with the initial state and ending in one of the states $s \in S$), and $S^{k+1}$ transition network CDAGs that depend on $k$ prior states, indexed by sequences of states visited, and ending in one of the states $s \in S$. Note that the initial transition networks all assume that the starting state is the initial state, while the transition network does not.

In addition, a $k$th-order Markov PDSEM makes the following assumption, that generalizes Assumption 2:

**Assumption 1** *For every state $s^j$, any CDAG $\mathcal{G}_{\langle\ldots j\ldots\rangle}$ or DAG $\mathcal{G}_j$ that mentions variables in state $j$ will have corresponding random variables that share state spaces.*

As was the case with DBNs, each DAG or CDAG in a PDSEM is associated with a causal model, which induces an appropriate DAG or CDAG factorization and g-formula for identification of interventional distributions. These, in turn, yield PDSEM factorizations that naturally generalize those in Section 4.2.

Let $\mathcal{T}_{\tilde{k}}$ be a set of all valid state transition sequences $\sigma_{\tilde{k}}$ of size $\tilde{k} = 1, \ldots, k-1$ that start with the initial state, and $\mathcal{T}$ be a set of all valid state transition sequences $\sigma_k$ of size $k$. Further, let $\mathbf{V}_\sigma$ be random variables in the final state in a state transition sequence $\sigma$, while $\mathbf{W}_\sigma$ be fixed variables in states prior to the final state in $\sigma$. Finally, let $\mathbb{I}(\sigma)$ be the indicator that the current state is the final state in $\sigma$, and the $|\sigma| - 1$ prior states were the states prior to the last state in $\sigma$. We then obtain the following observed data factorization of the $k$th-order Markov PDSEM:

$$p_1(\mathbf{V}_1) \prod_{t=1}^{k-1} \left( \prod_{\sigma_{\tilde{k}} \in \mathcal{T}_{\tilde{k}}} \left(p(\mathbf{V}_{\sigma_{\tilde{k}}} | \mathbf{W}_{\sigma_{\tilde{k}}})\right)^{\mathbb{I}(\sigma_{\tilde{k}})} \right) \prod_{t=k}^{\infty} \left( \prod_{\sigma_k \in \mathcal{T}} \left(p(\mathbf{V}_{\sigma_k} | \mathbf{W}_{\sigma_k})\right)^{\mathbb{I}(\sigma_k)} \right) 1^{\mathbb{I}(s_t^*)}$$

$$p_1(\mathbf{V}_1) = \prod_{V \in \mathbf{V}_1} p(V | \operatorname{pa}_{\mathcal{G}_1}(V));\; p_{\sigma_{\tilde{k}}}(\mathbf{V}_{\sigma_{\tilde{k}}} | \mathbf{W}_{\sigma_{\tilde{k}}}) = \prod_{V \in \mathbf{V}_{\sigma_{\tilde{k}}}} p(V | \operatorname{pa}_{\mathcal{G}_{\sigma_{\tilde{k}}}}(V));\; p_{\sigma_k}(\mathbf{V}_{\sigma_k} | \mathbf{W}_{\tilde{k}}) = \prod_{V \in \mathbf{V}_{\sigma_k}} p(V | \operatorname{pa}_{\mathcal{G}_{\sigma_k}}(V));$$

Extensions to truncated factorizations representing interventional distributions, and hidden variable versions of these models are straightforward generalizations of the $k = 1$ case, described in the main body of the paper.

As was the case with DBNs, relaxation of the first-order Markov assumption to a $k$th-order Markov assumption comes at a cost – many additional transition networks must be specified, and the resulting statistical inference is more likely to suffer from the curse of dimensionality.

### C.3 IDENTIFICATION IN CAUSAL DBNS AND PDSEMS THAT VIOLATE THE FIRST ORDER MARKOV ASSUMPTION

We show that identification in a hidden variable causal DBN for any finite horizon $T$ may be reformulated as a standard causal effect identification problem in a hidden variable causal model obtained by concatenating prior and transition networks in a DBN.

**Lemma 1** *Consider a causal DBN represented by the prior network DAG $\mathcal{G}_1$ with observed variables $\mathbf{V}_1$ and hidden variables $\mathbf{H}_1$, and the transition network conditional DAG $\mathcal{G}_{t+1,t}$ with observed variables $\mathbf{V}_{t+1}$, hidden variables $\mathbf{H}_{t+1}$, that depends on observed variables $\mathbf{V}_t$ and hidden variables $\mathbf{H}_t$ in the prior time step.*

*Given a fixed set of time points $1, \ldots, T$, let $\mathcal{G}_{1:T}$ be the unrolled hidden variable DAG obtained by concatenating the prior network $\mathcal{G}_1$, and the transition networks $\mathcal{G}_{t,t+1}$ for time points $2, \ldots, T$, $\mathbf{V}_{1:T} \equiv \bigcup_i \mathbf{V}_i$, and $\mathcal{G}_{1:T}(\mathbf{V}_{1:T})$ the latent projection of $\mathcal{G}_{1:T}$ onto $\mathbf{V}_{1:T}$.*

*Given disjoint $\mathbf{A}, \mathbf{Y} \subseteq \mathbf{V}_{1:T}$, $p(\mathbf{Y}(\mathbf{a}))$ is identified from $p(\mathbf{V}_{1:T})$ in the hidden variable causal DBN model represented by $\mathcal{G}_1$ and $\mathcal{G}_{t,t+1}$ if every district in $(\mathcal{G}_{1:T}(\mathbf{V}_{1:T}))_{\mathbf{Y}^*}$ (the induced subgraph of $\mathcal{G}_{1:T}(\mathbf{V}_{1:T})$) is intrinsic, where $\mathbf{Y}^*$ is the set of ancestors of $\mathbf{Y}$ in $\mathcal{G}_{1:T}(\mathbf{V}_{1:T})$ not through $\mathbf{A}$. Moreover, if $p(\mathbf{Y}(\mathbf{a}))$ is identified, it is equal to*

$$\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1:T}(\mathbf{V}_{1:T})_{\mathbf{Y}^*})} q_{\mathbf{D}}(\mathbf{D} | \operatorname{pa}_{\mathcal{G}_{1:T}}(\mathbf{D}) \setminus \mathbf{D})|_{\mathbf{A}=\mathbf{a}},$$

*where $q_{\mathbf{D}}$ are kernels corresponding to intrinsic sets that are districts in $\mathcal{D}(\mathcal{G}_{1:T}(\mathbf{V}_{1:T})_{\mathbf{Y}^*})$.*

*Proof:* This follows directly from standard results on identification in latent projections [Shpitser and Pearl, 2006, Richardson et al., 2017] and the fact that $\mathcal{G}_{1:T}$ represents a hidden variable causal model. □

Hence, if Assumption 1 does not hold, causal effects in causal DBNs are still identified for any finite number of timepoints $1, \ldots, T$, by constructing a hidden variable DAG via "unrolling" the hidden variable causal DBN for $T$ steps, applying the latent projection operation to this DAG to obtain an "unrolled" ADMG, and applying the ordinary ID algorithm to this ADMG. While this approach yields very general identification, the resulting functional will likely be computationally intractable, since the nested Markov factors in ADMGs that are not first-order Markov will generally be high dimensional objects.

We next show that, similarly to the result for causal DBN above, identification in a hidden variable PDSEM for any finite horizon up to time $T$ may be reformulated as a standard causal effect identification problem in a hidden variable causal model, provided the intervention on state variables $S_1, S_{ij}$ are part of the intervention set at every time point up to $T$.

Consider a PDSEM represented by the prior network DAG $\mathcal{G}_1$ with observed variables $\mathbf{V}_1$ and hidden variables $\mathbf{H}_1$, and a set of transition networks conditional DAGs $\mathcal{G}_{ij}$ (for any transition $(i.j) \in \mathcal{T})$) with observed variables $\mathbf{V}_{ij}$, hidden variables $\mathbf{H}_{ij}$, that depends on observed variables $\mathbf{V}_i$ and hidden variables $\mathbf{H}_i$ in the prior time step.

Given a fixed set of time points $1, \ldots, T$, and a particular sequence $\sigma \equiv \langle s_{k_1}, s_{k_2}, \ldots, s_{k_T} \rangle$ of states, such that $s_{k_1}$ is the initial state, and $(s_{k_i}, s_{k_{i+1}})$ is a valid state transition in $\mathcal{T}$, let $\mathcal{G}^{\sigma}_{1:T}$ be the unrolled hidden variable DAG obtained from $\mathcal{G}_1$ and $\{\mathcal{G}_{ij} : (i,j) \in \mathcal{T}\}$ inductively as follows. If $T = 1$, define $\mathcal{G}^{\sigma}_{1:T} \equiv \mathcal{G}_1$. Given $\mathcal{G}^{\sigma}_{1:i}$, define $\mathcal{G}^{\sigma}_{1:(i+1)}$ as follows: concatenate $\mathcal{G}_{s_i s_{i+1}}$ and $\mathcal{G}^{\sigma}_{1:i}$ by replacing every fixed (square) vertex in $\mathcal{G}_{s_i s_{i+1}}$ by the corresponding vertex in $\mathcal{G}^{\sigma}_{1:i}$ in time slice $i$.

Finally, define $V^{\sigma}_{1:T}$ as the set of all observed vertices in $\mathcal{G}^{\sigma}_{1:T}$, let $S^{\sigma}_{1:T}$ be the set of all state transition vertices in $\mathcal{G}^{\sigma}_{1:T}$, and define $\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})$ to be the latent projection of $\mathcal{G}^{\sigma}_{1:T}$ onto $\mathbf{V}_{1:T}$.

**Lemma 2** *Given disjoint $\mathbf{A}, \mathbf{Y} \subseteq \mathbf{V}_{1:T}$, such that $S^{\sigma}_{1:T} \subseteq \mathbf{A}$, and a value assignment $\mathbf{a}$ assigns $S^{\sigma}_{1:T}$ to the sequence $\langle s_{k_1}, \ldots, s_{k_T} \rangle$, $p(\mathbf{Y}(\mathbf{a}))$ is identified from $p_{\infty}(\mathbf{V})$ in the PDSEM model represented by $\mathcal{G}_1$ and $\{\mathcal{G}_{ij} : (i,j) \in \mathcal{T}\}$ if every district in $(\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T}))_{\mathbf{Y}^*}$ (the induced subgraph of $\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})$) is intrinsic, where $\mathbf{Y}^*$ is the set of ancestors of $\mathbf{Y}$ in $\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})$ not through $\mathbf{A}$. Moreover, if $p(\mathbf{Y}(\mathbf{a}))$ is identified, it is equal to*

$$\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})_{\mathbf{Y}^*})} q_{\mathbf{D}}(\mathbf{D} | \operatorname{pa}_{\mathcal{G}_{1:T}}(\mathbf{D}) \setminus \mathbf{D})|_{\mathbf{A}=\mathbf{a}},$$

*where $q_{\mathbf{D}}$ are kernels corresponding to intrinsic sets that are districts in $\mathcal{D}(\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})_{\mathbf{Y}^*})$, obtained by the usual sequential fixing operation applied to the distribution $p(\mathbf{V}_1) \cdot \prod_{t=1}^{T-1} p(\mathbf{V}_{s_{k_t}, s_{k_{t+1}}} | \mathbf{V}_{s_{k_t}})$ and $\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})$.*

*Proof:* This follows by induction on the length of the sequence $1:T$. The key observation is that since the treatment set $\mathbf{A}$ includes all state transition variables, the counterfactual evolution of the PDSEM under the intervention that sets $\mathbf{A}$ to $\mathbf{a}$, where state transition variables in $\mathbf{A}$ are set to values consistent with $\sigma = \langle s_{k_1}, s_{k_2}, \ldots, s_{k_T} \rangle$, is representable as a generalization of a causal DBN (where transition networks are not necessarily all equal, but where each transition network from $s_{k_i}$ to $s_{k_{i+1}}$ is obtained from the PDSEM). Further, the only part of the observed data distribution $p_{\infty}(\mathbf{V})$ of a PDSEM that is relevant to identification of the query are variables causally ancestral of the outcome in the graph containing all variables involved in transitions from 1 to $T$, which is just the graph $\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})$.

The result then follows directly from standard results on identification in latent projections [Shpitser and Pearl, 2006, Richardson et al., 2017], and the fact that $\mathcal{G}^{\sigma}_{1:T}(\mathbf{V}_{1:T})$ represents a hidden variable causal model. □

# D RELATED MODELING APPROACHES

## D.1 REPRESENTING A PDSEM AS A DBN

If variables in all transition networks in a PDSEM obey a single consistent topological order, one may encode a PDSEM by a causal DBN as follows. First, define a transition variable $T$ with values representing all possible state transition pairs $(s_i, s_j)$ in a PDSEM. Then, use this variable as a parent of every variable in the single transition network allowed by a DBN,

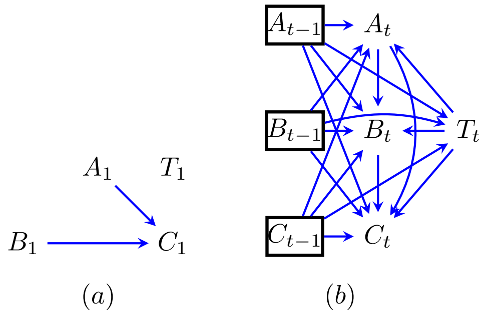


Figure 2: A causal DBN encoding the PDSEM in Fig. 4, via (a) the prior network, and (b) the complete transition network with context-specific independences.

and use it to select a subset of all possible parents to implement transition specific networks of a PDSEM via context-specific independence.

In the example in the Section 4.1 shown in Fig. 4, one topological order on variables that is consistent for the prior network and all transition networks is $A \prec B \prec C$. Thus, a causal DBN representing the example PDSEM would have a prior network shown in Fig. 2 (a), and a complete conditional DAG as a transition network shown in Fig. 2 (b), with a factorization: $p(C_t|B_t, A_t, T_t, V_{t-1})) \cdot p(B_t|A_t, T_t, V_{t-1}) \cdot p(A_t|T_t, V_{t-1}) \cdot p(T_t|V_{t-1})$, where $V_{t-1} \equiv C_{t-1}, B_{t-1}, A_{t-1}, T_{t-1}$. Note that in this transition network, every state variable has the transition variable as a parent, and this parent is used to implement state transition independences in a PDSEM via context-specific independence. For example, the Markov factor $p(B_t|A_t, T_t, C_{t-1}, B_{t-1}, A_{t-1}, T_{t-1})$ will not depend on $A_t$ unless $T_t$ has value $(s_2, s_3)$.

Note that this representation, is in some sense, isomorphic to PDSEMs. The causal DBN factorization exhibits no independences, and all interesting probabilistic and causal structure is obtained via context-specific independences, which would be represented explicitly in transition networks of a PDSEM.

In addition, if no consistent topological order on variables in all transition networks in a PDSEM exists, then there is no known representation scheme for such a PDSEM using causal DBNs.

### D.2 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 $\big(\mathcal{V}, \mathcal{A}, \mathcal{R}, p(V_{t+1} = v', R_{t+1} = r | V_t = v, A_t = a), \gamma\big)$ where $\mathcal{V}$ is a finite set of states, $\mathcal{A}$ is a set of actions, $\mathcal{R}$ 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 $\pi(a|v) : \mathcal{V} \mapsto \mathcal{A}$ 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:

$$G_\pi(v) = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid V_t = v\Big] = \sum_a \pi(a|v) \sum_{s'} \sum_r p(s', r|s, a)\Big[r + \gamma G_\pi(v')\Big] \tag{4}$$

This can be viewed as a consistency condition between values $G_\pi(v)$ and $G_\pi(v')$ 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 $v \in \mathcal{V}$. The optimal policy $\pi^*$ may not be unique, and has optimal value function: $G^*(v) = \max_\pi G_\pi(v)$ for all $v \in \mathcal{V}$.

**An important special case:** Consider an MDP with the following features: First, there are absorbing states $V^* \equiv \{v_1^*, \ldots v_k^*\}$. 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 $v_t \notin V^*$ or if $v_{t-1} \in V^*$. And $R_t = r(v_i^*)$ if $v_t \in V^*$ and $v_{t-1} \notin V^*$. 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:

$$G(v) = \sum_{V^*=v_i^*} r(v_i^*) \sum_{k=1}^{\infty} p^k(v_i^*|v) = \sum_{V^*=v_i^*} r(v_i^*) p^{\infty}(v_i^*|v), \tag{5}$$

where $p^k(v_*^i|v)$ is the $k$-step transition probability from $v$ to $v_i^*$, and $p^{\infty}(v_i^*|v)$ is the probability of eventually reaching $v_i^*$ 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 $V_0 \equiv \{v_{01}, v_{02}, v_{03}\}$, $V_1 \equiv \{v_{11}, v_{12}, v_{13}\}$ and $V_2 \equiv V^* \equiv \{v_1^*, v_2^*, v_3^*\}$. We have the following simple transition diagram: $v_{0i} \to v_{1j} \to v_k^*$, for all $i, j, k$ and the above expression for total expected reward yields:

$$\sum_{V_0, V_1, V^*} r(V^*) p(V^*|V_1, a_0) p(V_1|V_0, a_0) p(V_0) = \mathbb{E}_{q_a}[r(V^*)], \tag{6}$$

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

$$\sum_{V_0, V_1, V^*} r(V^*) p(V^*|V_1, a_1 = \pi(V_1)) p(V_1|V_0, a_0) = \pi(V_0)) p(V_0) = \mathbb{E}_{q_\pi}[r(V^*)], \tag{7}$$

Equations (6) and (7) 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. [Zhang and Bareinboim, 2016] introduce Markov Decision Processes with Unobserved Confounders (MDPUCs) that are (causal) MDPs with local unobserved variables counfounding relationships between actions, effects and states, while obeying Markovian dynamics (i.e., all previous states and actions can be best summarized by the current state). There is more work related to causal MDPs and learning associated policies in the causal reinforcement literature [Forney et al., 2017, Zhang and Bareinboim, 2019, Zhang, 2020]. Our work on PDSEMs generalizes both causal DBNs and causal MDPs, by capturing complex state dynamics and complex state transitions in the presence of unobserved confounding.

## E PROOFS

**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 representing elements of $\mathcal{D}(\mathcal{G}_{1\mathbf{Y}_1^*})$ 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.*

*Proof:* We want to obtain $p(\mathbf{Y}(\mathbf{a}))$ from the observed joint $p(\mathbf{V}_{1:T})$. Using identification result 5 on the unrolled ADMG gives $\sum_{Y^* \setminus Y} p(\mathbf{Y}^*(a)) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{D \in \mathcal{D}(\mathcal{G}_{\text{unrolled}\mathbf{Y}^*})} p(\mathbf{D}(\operatorname{pa}(\mathbf{D}) \setminus \mathbf{D}))|_{\mathbf{A}=\mathbf{a}}$. Assumption 1 ensures that no district $\mathbf{D}$ spans time points, and parents $\operatorname{pa}(\mathbf{D})$ at time $t+1$ lie either at $t$ or $t+1$. This allows us to write $\sum_{\mathbf{Y}^* \setminus \mathbf{Y}} p(\mathbf{Y}^*(\mathbf{a})) = \sum_{\mathbf{Y}^* \setminus \mathbf{Y}} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{1Y^*})} p(\mathbf{D}(\operatorname{pa}(\mathbf{D}) \setminus \mathbf{D}))|_{\mathbf{A}=\mathbf{a}} \times \prod_{t=1}^{T-1} \prod_{\mathbf{D} \in \mathcal{D}(\mathcal{G}_{t+1,tY^*})} p(\mathbf{D}(\operatorname{pa}(\mathbf{D}) \setminus \mathbf{D}))|_{\mathbf{A}=\mathbf{a}}$. Applying the identification results in [Richardson et al., 2012] to the prior network ADMG $\mathcal{G}_1$ and extensions of these results in [Sherman and Shpitser, 2018] to the transition network CADMGs $\mathcal{G}_{t+1,t}$, these counterfactual conditionals can be replaced by given modified nested factorizations, provided every appropriate bidirected connected set in the prior or transition graph is intrinsic in that graph.

Note that completeness of our procedure does not immediately follow from the completeness argument in [Shpitser and Pearl, 2006]. This is because a completeness argument entails constructing in any ADMG $\mathcal{G}(\mathbf{V})$ where identification fails two causal models which agree on the observed data distribution $p(\mathbf{V})$, but disagree on $p(\mathbf{Y}(\mathbf{a}))$. Furthermore, the construction employed in [Shpitser and Pearl, 2006] relied on an unrestricted causal model inducing a given latent projection ADMG

$\mathcal{G}(\mathbf{V})$. However, in the case of causal DBNs, the model is *not* unrestricted – indeed there is a very strong restriction that all transition networks at any time point share all structural equations.

Nevertheless, it is possible to extend the completeness proof in [Shpitser and Pearl, 2006] to yield completeness of the procedure in this lemma by employing an extended construction modeled after one in [Shpitser and Sherman, 2018].

From this point on, we will refer to $\mathcal{G}_{1:T}(\mathbf{V}_{1:T})$ by $\mathcal{G}(\mathbf{V})$ for simplicity. Assume $p(\mathbf{Y}(\mathbf{a}))$ is not identified in $\mathcal{G}(\mathbf{V})$, and assume there exists a hedge structure ancestral of $\mathbf{Y}'$. Note that by first order Markov assumption, the hedge structure must lie entirely in a transition network in a single time step. Fix a subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}(\mathbf{V})$ containing the hedge, the set $\mathbf{Y}$, a set of vertices $\mathbf{S}$ making up directed paths from every element of the root set $\mathbf{R}$ to some element of $\mathbf{Y}'$ (without loss of generality we assume these vertices do not have more than one child).

We extend $\tilde{\mathcal{G}}$ with a new set of vertices $\mathbf{S}^*$ that are copies of $\mathbf{S}$ with the property that if $S \in \mathbf{S}$ has a parent in $\mathbf{R}$, so does the corresponding $S^* \in \mathbf{S}^*$, and if $T \in \mathbf{S}$ is a parent of $S \in \mathbf{S}$, the corresponding $T^* \in \mathbf{S}^*$ is a parent of $S^* \in \mathbf{S}^*$. We then apply the counterexample construction connecting the hedge structure to $\mathbf{Y}'$ appearing in [Shpitser and Pearl, 2006] to elements of $\mathbf{S}^*$. In particular, we make sure that $\sum_{\mathbf{S}^*} p(\mathbf{Y}'|\mathbf{S}^*)p(\mathbf{S}^*|\mathbf{R})$ is a one-to-one map. This implies $p(\mathbf{Y}(\mathbf{a}))$ is not identified in an extended model containing vertices $\mathbf{V}$ and $\mathbf{S}^*$. Lemma 1 in [Shpitser and Sherman, 2018] then implies $p(\mathbf{Y}(\mathbf{a}))$ is also not identified in $\mathcal{G}(\mathbf{V})$, establishing our result. □

**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:*

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

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

*Proof:* This follows from the factorization of $p_\infty(\mathbf{V}(\mathbf{a}))$ into elements of the form $p_1(\mathbf{Y}_1(\mathbf{a}_1))$, and $p_{ij}(\mathbf{Y}_j(\mathbf{a}_j)|\mathbf{Y}_i(\mathbf{a_i}))$, the fact that $\mathcal{G}_1, \{\mathcal{G}_{ij} : (i,j) \in \mathcal{T}\}$ define causal models under standard structural equation semantics, and equation 1 . □

**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*

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

*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 $\mathcal{G}_{ij}$.*

*Proof:* Assumption 3 implies all state transitions are known, and thus allows us to proceed by induction on any sequence of state transitions with positive probability after $t$ steps.

Unrolling the prior network, and appropriate transition networks for such a sequence yields an ADMG representing the observed data distribution had that transition taken place, with Assumption 1 implying that districts in this ADMG do not span multiple time steps. This immediately implies the conclusion by the same argument used in the proof of Lemma 1.

In fact, this argument works for any transition sequence of any size. □

# F 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 often divided into discrete steps or "stages", each with its own intermediate goal [Ahmidi et al., 2015]. Our data from the septoplasty procedure was manually annotated by clinical experts and divided into the following states:

- $s_1$: 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: $\mathbf{V} = \{$ K: knife, G: gorney scissors, $C_1$: cottle, $D_1$: short needle driver, $D_2$: long needle driver, O: other tools, $C_2$: suction cannula, M: main surgeon exists, S: suction exists, $A_1$: main surgeon is an attending, $A_2$: suction done by attending, T: duration of that phase is greater than 10 seconds $\}$

- $\mathbf{V}_{s_1} = \{\text{K, O, C}_2\text{, M, S, A}_1\text{, A}_2\text{, T}\}$,
- $\mathbf{V}_{s_2} = \{\text{K, C}_1\text{, O, C}_2\text{, M, S, A}_1\text{, A}_2\text{, T}\}$,
- $\mathbf{V}_{s_3} = \{\text{K, C}_1\text{, D}_1\text{, D}_2\text{, O, C}_2\text{, M, S, A}_1\text{, A}_2\text{, T}\}$,
- $\mathbf{V}_{s_4} = \{\text{K, C}_1\text{, G, O, C}_2\text{, M, S, A}_1\text{, A}_2\text{, T}\}$,
- $\mathbf{V}_{s_5} = \{\text{D}_1\text{, D}_2\text{, O, C}_2\text{, M, S, A}_1\text{, A}_2\text{, T}\}$,
- $\mathbf{V}_{s_6} = \{\text{K, C}_1\text{, O, C}_2\text{, M, S, A}_1\text{, A}_2\text{, T}\}$,

Since our target of interest is causal effect of surgeon experience on average surgery length, the interventions are considered on variables $A_1$ and $A_2$.

To determine the allowed state transitions , we retained observed data state transitions where at least 5 such transitions occurred. The permitted state transitions $s_i \to s_j$ are summarized in Figure 7 in the main paper – note that transitions other than those depicted have probability $p(s_j \mid s_i, \mathbf{v}_{s_i}) = 0$ for all $\mathbf{v}_{s_i}$. To determine the state transition distributions $p(s_j \mid s_i, \mathbf{v}_{s_i})$, we restricted the set $\mathbf{v}_{s_i}$ for all $i$ to be $\{A_1, M, A_2, S\}$ to increase tractability of estimation, and estimated this discrete conditional distribution via a conditional probability table. The prior distribution on the initial state was set to $p(s_1) = 1$.

State DAGs were determined based on clinician recommendation and have been reproduced in Figure 3 for reference. These immediately lead to prior variable distributions $p(\mathbf{V}_{s_i})$ for each state $s_i$.

Transition graphs from $s_i \to s_j$ are constructed using a simple rule: the for any variable $v$ in any state $s_j$, the parents $\text{pa}(v_{s_j})$ consists of the variable with the same name in the previous state $s_i$ if it exists, and all parents in the state DAG for $s_j$ point indicated by state DAGs. For example, in the transition $s_1 \to s_1$ moving from time step $t-1 \to t$, variable $K$ at time step $t$ has parents $A_1$ at $t$, as given in Figure 3(a), as well as $K$ from time step $t-1$. However, in the transition $s_1 \to s_2$ moving from time step $t-1$ to time step $t$, variable $C_1$ has parent $A_1$ in time step $t$, as given in Figure 3(b), but no parents from the previous time step $t-1$ since $C_1$ does not exist in $s_1$. Based on this rule, probability distributions $p(v_{s_j} \mid \text{pa}(v_{s_j}))$ are estimated using conditional probability tables.

Goodness of fit of our model with respect to the original data distribution is shown in Figure 4. Trajectories simulated by our model are able to capture the distribution of surgery duration originally seen in the data, quite well.

### F.1 STATISTICAL INFERENCE

Given counterfactual distributions identified via (7), (9), and (10) in the main paper, if a parametric likelihood may be specified in terms of components of these equations, statistical inference may be performed by plug-in estimation and

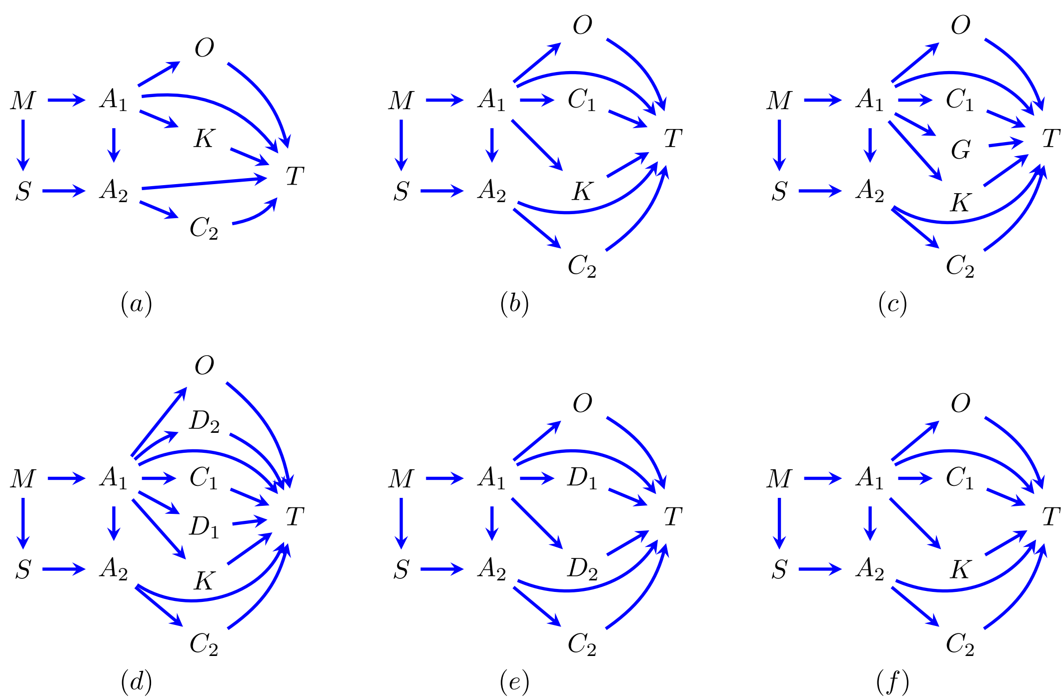


Figure 3: (a)-(f) State DAGs corresponding to states $s^1$: Opening of the septum, $s^2$: Raising septal flaps, $s^3$: Removal of deviated septal cartilage and bone, $s^4$: Reconstruction, $s^5$: Closing of incision and $s^6$: Activity not otherwise included in the above phases.

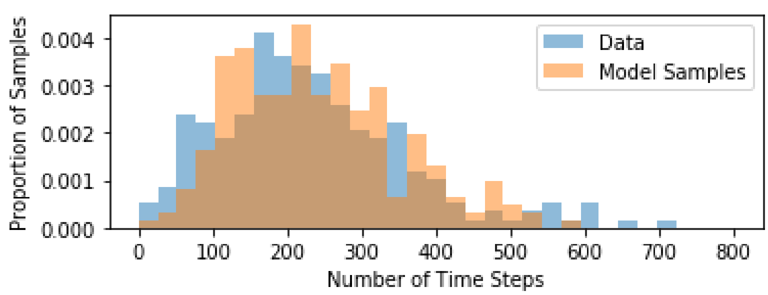


Figure 4: Histograms of observed surgery (blue) versus simulated surgeries from the estimated model (orange).

Monte-Carlo simulation. In a fully observed PDSEM, the likelihood may be obtained from (6) by imposing parametric models for every Markov factor. In a hidden variable PDSEM, the likelihood may be obtained from the nested Markov factorization for the marginal distribution associated with the prior network ADMG, and the conditional distributions associated with transition network CADMGs. These likelihoods are available in multivariate normal assumption on the observed data, which we illustrate via the simulation study, and discrete state spaces, via the Moebius inversion formula parameterization discussed in [Evans and Richardson, 2018].

Given model parameters obtained by maximizing the PDSEM observed data likelihood, counterfactual distributions in (7), (9), and (10) (of the main paper) may be obtained by simulating PDSEM trajectories using these modified factorizations, evaluated at MLE parameter values. Confidence intervals for any counterfactual parameter of interest may be obtained by parametric bootstrap.

However, an analogous approach is not straightforward for nested Markov parameterizations of the marginal PDSEM representing a PDSEM with hidden variables. In our simulations, we use a specific generative model for our continuous variables, i.e, the linear Gaussian Structural Equation model. Another choice based on work in [Evans and Richardson, 2014] is the Möbius parameterization for binary variables. However, this 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 the joint distribution $p(\mathbf{V})$ (from which it is easy to generate samples for a discrete sample space of $\mathbf{V}$). Importantly, such a conversion leads to an intractable object that requires storage and running time exponential in $|\mathbf{V}|$. 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. For settings beyond Gaussian and discrete data, statistical inference strategies are significantly more complicated and have been discussed in [Bhattacharya et al., 2020].

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.

# G COMPUTATION DETAILS

The septoplasty data application presented in Section 6 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 each.

### 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.

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

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

Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *Annals of Statistics*, pages 1–30, 2014.

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.

Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In *International Conference on Machine Learning*, pages 1156–1164. PMLR, 2017.

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

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.

Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. In *Advances in neural information processing systems*, pages 9424–9435, 2018.

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.

Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

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, 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.

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.

Junzhe Zhang. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR, 2020.

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

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. 2019.