Do-calculus enables causal reasoning with latent variable models

Preprint, compiled February 15, 2021

Sara Mohammad-Taheri[1], **Robert Ness**[2], **Jeremy Zucker**[3], **and Olga Vitek**[1*]

[1]Khoury college of computer sciences, Northeastern University, Boston, MA, USA
[2]Altdeep, Boston, MA, USA
[3]Pacific Northwest National Laboratory, Richland, WA

Publication doi

Abstract

Latent variable models (LVMs) are probabilistic models where some of the variables are hidden during training. A broad class of LVMshave a directed acyclic graphical structure. The directed structure suggests an intuitive causal explanation of the data generating process. For example, a latent topic model suggests that topics cause the occurrence of a token. Despite this intuitive causal interpretation, a directed acyclic latent variable model trained on data is generally insufficient for causal reasoning, as the required model parameters may not be uniquely identified. In this manuscript we demonstrate that an LVM can answer any causal query posed post-training, provided that the query can be identified from the observed variables according to the do-calculus rules. We show that causal reasoning can enhance a broad class of LVM long established in the probabilistic modeling community, and demonstrate its effectiveness on several case studies. These include a machine learning model with multiple causes where there exists a set of latent confounders and a mediator between the causes and the outcome variable, a study where the identifiable causal query cannot be estimated using the front-door or back-door criterion, a case study that captures unobserved crosstalk between two biological signaling pathways, and a COVID-19 expert system that identifies multiple causal queries.

1 Introduction

Latent variable models (LVMs) are probabilistic models of a joint distribution on a set of variables, where some of the variables are unobserved (i.e., hidden, latent) during training. These models have had a major impact on applications ranging from natural language processing, social science, computer vision, to computational biology. A broad class of LVMs have a directed acyclic graphical structure. Canonical examples of directed acyclic graphical LVMs include topic models, hidden Markov models, Gaussian mixture models [1], and deep generative latent variable models such as variational autoencoders [2].

LVMs contain parameters that must be learned from data. Once the model is trained, it can answer numerous marginal and conditional queries with respect to its variables using computational inference algorithms, including graphical modeling inference algorithms such as belief propagation [3], gradient-based sampling techniques such as Hamiltonian Monte Carlo [4] or stochastic variational inference [5]. The ability to answer multiple queries after a single training is particularly valuable for probabilistic reasoning systems, e.g. in natural language processing or medical diagnosis, where large-scale models are expensive to train and maintain. When variables are latent during training, some model parameters and the associated model-based queries may not be uniquely identified.

The directed graph structure of latent variable models often suggests intuitive causal interpretations. Indeed, for many LVMs the causal semantics is built into the model definition. For example in natural language processing, the directed structure of a latent topic model suggests that topics drive the occurrence of token. In molecular biology, the directed structure may suggest a process by which an interaction between a virus and a host protein dysregulates the immune response, even in the absence of measurements of that interaction. Formalizing that causal intuition into the ability to answer causal queries would greatly extend the power of LVMs. An LVM that could answer *ad hoc* causal queries would be advantageous as compared to alternative methods of causal inference, e.g. the "plug-in" estimator [6], that require building a new statistical model every time a new causal query arises. Unfortunately, due to the non-uniqueness of parameters during training, LVM-based estimates of causal queries are in general incorrect.

This manuscript proposes an approach for alleviating the challenge above. Using Bayesian perspective, we demonstrate that training an LVM, and applying probabilistic inference algorithms to the trained model yields an accurate estimator of a causal estimand, provided that the estimand is identifiable from the training variables and the graph structure according to a set of rules called the do-calculus.

We illustrate the generality and the practical utility of causal reasoning with LVMs in four case studies. The first is an LVM with multiple causes, similar to the Box Office revenue model in [7], modified to include a mediator [8, 9]. Such structure underlies many high-dimensional problems in generative machine learning. The second case study is a deceptively simple causal LVM, known as the new Napkin problem [10]. It shows the applicability of the proposed approach to a graph topology where causal reasoning is challenging. The third case study is an LVM that captures unobserved crosstalk between two signaling pathways. The fourth case study uses a molecular biology expert system to model host response to viral infection of SARS-CoV-2 the novel coronavirus responsible for the COVID-19 pandemic. In this case, multiple causal queries were identified by introducing latent variables that isolate each causal effect from the rest of the system.

*correspondence: ovitek@northeastern.edu

2 BACKGROUND

2.1 Notation

Let bold face upper case letters such as $\mathbf{X} = \{X_1, ..., X_J\}$ be a set of random variables and an upper case letter such as X_i be a single variable. Note that X_i can be multivariate. Let $\mathbf{x}$ be an instance of $\mathbf{X}$ and x_i be an instance of X_i. Let $P(x_1, ..., x_j)$ be the joint probability distribution of the event $\mathbf{X} = \mathbf{x} = \{x_1, ..., x_j\}$. $P(X_i = x_i | X_j = x_j)$ is a conditional probability distribution for the event $X_i = x_i$ given $X_j = x_j$. In this manuscript, we simplify the joint distribution as $P(\mathbf{x})$, and conditional distribution as $P(x_i|x_j)$. Let G be a directed acyclic graph (DAG) with nodes $\mathbf{X}$, where $\text{Pa}(X_i)$ are the parents of a node X_i in G. The causal Markov condition assumes that $P(\mathbf{x})$ factorizes along the structure of G, i.e. $P(\mathbf{x}) = \prod_{x_i \in \mathbf{X}} P(x_i|\text{Pa}(x_i))$.

2.2 Latent variable models

Latent variable models (LVM) are probabilistic models of $P(\mathbf{x})$, where some variables are not observed during training. LVMs are generative, in the sense that they allow us to generate samples from $P(\mathbf{x})$. A particularly attractive class are LVM with directed acyclic graph structures. Beyond representing conditional independence on $P(\mathbf{x})$, the structures typically have an intuitive causal interpretation. In this manuscript we refer to latent variable models with DAG structures and causal interpretation as causal LVM.

A **latent mediator** is a latent variable with incoming and outgoing edges. A node X_k is a **collider** if it is part of a $X_i \leftarrow X_k \rightarrow X_j$ structure with no edges connecting X_i and X_j. A **latent confounder** is a latent node affecting both the cause and the effect. After the training, a causal LVM is a probabilistic expert system that can answer many causal and non-causal queries about variables in $\mathbf{X}$ by applying probabilistic inference algorithms such as MCMC and stochastic variational inference (SVI).

2.3 Causal inference

An **intervention** on a target variable X fixes the variable to a constant x (denoted $do(x)$ [8]), rendering it independent of its causes [11, 12]. The **causal effect** of X on Y is denoted $P(y|do(x))$.

Graph mutilation in a causal graphical model is a method for simulating the effect of an intervention. Graph mutilation severs the incoming edges to the target node, and fixes it to the intervention value [13] rendering it independent of its causes. In the following we denote $G_{\bar{X}}$ the graph produced by mutilating G to remove all incoming edges to X, and $P_{G_{\bar{X}}}(x)$ denotes the new distribution created by the mutilation. Sampling from $P(y|do(x))$ is achieved by applying algorithmic inference to $G_{\bar{X}}$ and sampling from $P_{G_{\bar{X}}}(y|x)$.

A **causal query** is any probabilistic query that conditions on an intervention, such as $P(x_j|do(x_i), x_k)$ or $E[x_j|do(x_i)]$. It is answered either by applying inference algorithms to the mutilated graph representing the joint interventional distribution, or by estimating the equivalent probabilistic expression on the joint observational distribution. In other work, the term causal query includes counterfactual queries [8]. While many counterfactual queries reduce to conditioning on interventions, we consider general counterfactual queries beyond the scope of this work.

The **average causal effect** of X on Y is a special case of causal query defined as $E[Y|do(x)] - E[Y]$.

A causal query on a joint interventional distribution is **identifiable** if it can be transformed into an equivalent probabilistic expression on the joint observational distribution. This can be determined by the following criteria.

The **back-door criterion** holds as long as there are no unobserved confounders of cause and effect. If the back-door criterion holds, $P(Y|do(x'))$ is identifiable and is given by $\int_{\mathbf{z}} P(y|x', \mathbf{z})P(\mathbf{z})dz$ where $\mathbf{Z}$ is a set of variables that satisfies the back-door criterion relative to X and an effect Y in a DAG G.

The **front-door criterion** holds even when there is an unobserved confounder, but there exists a mediator (M) between cause and effect that is shielded from confounding. If the front-door criterion holds, $P(Y|do(x'))$ is identifiable and is given by, $\int_m \left(\int_x P_G(Y|m, x)P_G(x)dx \right) P_G(m|x')dm$ [8, 9]. Front-door criterion is particularly useful when the back-door criterion does not hold. For example, the back-door criterion does not hold in Fig. 1 (a) and (b) but the front-door criterion does hold in Fig. 1 (b).

The back-door and front-door criteria are sufficient but not necessary for causal identifiability. The **do-calculus**, comprised of three graph-mutilation-based rules [14], helps determine other identifiable causal queries. A causal query containing a $do()$ operator is identifiable if the do-calculus transforms it into an equivalent do-free estimand. The do-calculus estimands are *non-parametric* in the sense that they do not impose constraints on $P(\mathbf{x})$.

Depending on the causal query, the graph, and the set of latent variables, a number of sound and complete algorithms and implementations help determine whether the query is identifiable [15, 16], and, if the query is identifiable, generate an estimand [17, 15, 18, 19].

2.4 Expressing model misspecification

Model misspecification is a common challenge in modeling joint probability distributions, particularly when some variables are latent. Acyclic directed mixed graph (ADMG) [16] structures help account for potential model misspecification. ADMG ignore latent mediators and colliders, and only consider the effect of latent confounding. The presence of latent confounders in an ADMG is indicated with bidirected ($\leftrightarrow$) edges.

An ADMG has several properties. First, it is associated with a model that has the same equality constraints (such as conditional independence relationships) on the observed joint probability distribution as those obtained by marginalizing out the latent variables from the original LVM [20]. At the same time, the model associated with the ADMG does not contain inequality constraints (such as nonparametric bounds on instrumental variables) that may exist in the original LVM [21, 22]. Second, an infinite set of causal LVMs can project onto the same ADMG. Third, any causal query in an ADMG identifiable by the do-calculus rules is also identifiable in every causal LVM that projects onto that ADMG [16]

2.5 *Existing methods for inferring causal effects*

Developing estimators of causal effects of a set of variables $\mathbf{X}$ on a variable Y has been subject of much research. E.g., the plug-in estimator [6] is a straightforward approach, that assumes a parametric model for the conditional distributions appearing in a do-calculus-based estimand. If the formula for calculating the estimand is complicated and includes many conditional distributions, one has to make parametric assumptions for each distribution. Other authors [23, 7, 24] have developed generative machine learning modeling techniques by relying on the presence of multiple causes or the proxy variables.

Many algorithms efficiently estimate the causal effect when the back-door criterion holds [25, 26, 27]. Famous examples include the g-formula [28], the inverse probability weighting (IPW) [29], and Targeted Maximum Likelihood Estimators (TMLE), a family of doubly-robust estimators combining the g-formula and IPW. Augmented IPW [30, 31] is a doubly robust semiparametric approach that goes beyond back-door criterion in specific settings. For models with a single cause and a single outcome, [32] propose doubly robust semiparametric estimators primal IPW (PIPW) and dual IPW (DIPW) that do not require the back-door criterion to hold.

[33] proposed weighted empirical risk minimization (WERM-ID), a general probabilistic inference technique for any do-calculus-identified estimand. Unfortunately at the time of writing its implementation was not publicly available for evaluation.

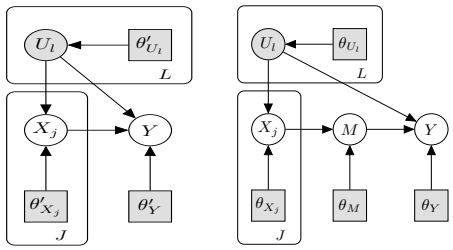
For problems with a network structure, data, and an identifiable causal query, the `ananke` library [34, 35] suggests appropriate estimators (including IPW, AIPW, g-formula, PIPW and DIPW) and the corresponding estimate. The Causal Fusion platform [19] is another tool that takes as input a DAG and a causal query, and investigates the identifiability of the query. If the query is identifiable, it provides a full derivation of the do-calculus-based formula.

The methods above start with a pre-specified causal query, derive an estimand for that query based on theoretical premises such as do-calculus, and derive an estimator for that query. The process repeats every time the new queries arise. In contrast, this manuscript targets a workflow that trains a causal causal latent variable model once, and answers multiple causal queries by applying graph mutilation and algorithmic inference to the trained model. We propose an approach that does not rely on the back-door or the front-door criteria, is applicable to multiple causes and outcomes, and restricts parametric assumptions to the data-generating process.

3 METHODS

In this section we show that training a causal latent variable model (LVM) then applying graphical inference on the trained mutilated model is equivalent to estimating a do-calculus identifiable estimand. Taking a Bayesian view, we first demonstrate this in the special case of Fig. 1(b), where the causal effect of $\mathbf{X}$ on Y is identifiable according to the front-door criterion. Then we show that this is true for any identifiable causal query in an arbitrary causal LVM.

Lemma *Assume the parameters of an LVM are the ground truth parameters. Then exact graph-based probabilistic inference of a causal query on the mutilated LVM is an estimand identified by the do-calculus.*



(a) (b)

Figure 1: Plate representation of LVM. Circular white/gray nodes are observed/latent variables. Square gray nodes are the associated parameters. Each parameter such as θ_U has a prior distribution, e.g. $\theta_U \sim P(q_{\theta_U})$, where q_{θ_U} is a hyperparameter. Each variable is conditionally independent of its non-descendents given its parents. (a) Multi-cause model without mediator. $P(Y|do(\mathbf{x}))$ is not non-parametrically identified. (b) Multi-cause model with a mediator. $P(Y|do(\mathbf{x}))$ is non-parametrically identified.

Proof. According to the do-calculus, a causal query involving an intervention on X transforms a probability distribution encoded by $P_G(.)$, to a distribution encoded by $P_{G_{\bar{\mathbf{x}}}}(.)$. Then, given a ground truth parameterization of $P_G(.)$, exact inference on $P_{G_{\bar{\mathbf{x}}}}(.)$ samples from the query distribution.

Assume a causal DAG G as in Fig. 1(b), where $\mathbf{X}$, M, and Y are observed variables, and $\mathbf{U}$ is latent, and the query $P_G(Y|do(\mathbf{x}'))$. Modeling ideal interventions with graph mutilation implies that $P_G(Y|do(\mathbf{x}')) = P_{G_{\bar{\mathbf{x}}}}(Y|\mathbf{x}')$. Hence,

$$
\begin{aligned}
& P_G(Y|do(\mathbf{x}')) = P_{G_{\bar{x}}}(Y|\mathbf{x}') \\
&= \int_{\mathbf{u},m} P_{G_{\bar{\mathbf{x}}}}(Y,\mathbf{u},m|\mathbf{x}')dmd\mathbf{u} \\
&= \int_m \left(\int_{\mathbf{u}} P_{G_{\bar{\mathbf{x}}}}(Y|\mathbf{u},m,\mathbf{x}')P_{G_{\bar{\mathbf{x}}}}(\mathbf{u}|m,\mathbf{x}')d\mathbf{u} \right) P_{G_{\bar{\mathbf{x}}}}(m|\mathbf{x}')dm \\
&= \int_m P_{G_{\bar{\mathbf{x}}}}(Y|m,\mathbf{x}')P_{G_{\bar{\mathbf{x}}}}(m|\mathbf{x}')dm \\
&= \int_m P_{G_{\bar{\mathbf{x}}}}(Y|m)P_G(m|\mathbf{x}')dm \qquad (1) \\
&= \int_m P_G(Y|do(m))P_G(m|\mathbf{x}')dm \\
&= \int_m \left(\int_{\mathbf{x}} P_G(Y|m,\mathbf{x})P_G(\mathbf{x})d\mathbf{x} \right) P_G(m|\mathbf{x}')dm \qquad (2)
\end{aligned}
$$

Eq. (1) holds because Y is independent from $\mathbf{x}$ given m in $G_{\bar{\mathbf{x}}}$. $P_{G_{\bar{\mathbf{x}}}}(m|\mathbf{x}')$ is unaffected by the mutilation of G that creates $G_{\bar{\mathbf{x}}}$. Hence, $P_{G_{\bar{\mathbf{x}}}}(m|\mathbf{x}') = P_G(m|\mathbf{x}')$. Eq. (2) follows from the back-door path between Y and M in G. The expression on the right-hand side of Eq. (2) is exactly the front-door adjustment formula, an estimand for $P_G(Y|do(\mathbf{x}'))$ that is derived from the do-calculus. □

The lemma illustrates that if a causal query on an LVM is identified by the do-calculus given the variables observed during training, then exact inference on the mutilated LVM is a valid

do-calculus estimand. Next we show that combining this procedure with the training step constitutes an estimator for that estimand. While we chose to prove this theorem using concepts common to Bayesian statistics, we also provide an equivalent proof using concepts from causal inference in the supplementary materials.

Theorem *Assume that a causal query of the LVM is identifiable via do-calculus given the set of variables observed during training. Then the training procedure combined with exact sampling-based inference on the mutilated model yield an estimator of that query that converges to the unique true value as the number of samples goes to infinity.*

Proof. Assume a causal DAG $G(\mathbf{X})$, where $\mathbf{X}_o \subseteq \mathbf{X}$ are observed during training. Let $P_G(.)$ be a probability distribution encoded by G.

Both the model parameters and the latent variables are the subjects of inference during training. Let Θ be a random vector representing the union of the model parameters θ and of the latent variables. Inference on a causal query depends on Θ. Let $Q(\Theta)$ denote the causal query. We infer its expectation via the posterior inference over Θ:

$$E(Q(\Theta)) = \int_{\Theta} Q(\Theta) P_G(\Theta|x_o) d\Theta$$

When the model is trained, some elements in Θ are identified (i.e., inference converges to a unique solution as the size of the data goes to infinity), and some are not. Partition $\Theta = \{\phi, \lambda\}$, where ϕ contains the identified components and λ are unidentified. If the query is identified by the do-calculus, then by definition any estimand of $Q(\Theta)$ is a function only of $\mathbf{x}_o$, i.e. any valid estimator for an estimand for $Q(\Theta)$ can only rely on parameters uniquely identified by $\mathbf{x}_o$. Therefore, $Q(\Theta) = Q(\phi)$, i.e. inference of $Q(\Theta)$ only depends on the parameters in ϕ.

$$\begin{aligned}
E(Q(\Theta)) &= \int_{\Theta} Q(\phi) P_G(\Theta|x_o) d\Theta \\
&= \int_{\phi,\lambda} Q(\phi) P_G(\lambda, \phi|x_o) d\phi, d\lambda \\
&= \int_{\phi,\lambda} Q(\phi) P_G(\lambda|\phi, x_o) P_G(\phi|x_o) d\phi, d\lambda \\
&= \int_{\phi} Q(\phi) \left(\int_{\lambda} P_G(\lambda|\phi, x_o) d\lambda \right) P_G(\phi|x_o) d\phi \\
&= \int_{\phi} Q(\phi) P_G(\phi|x_o) d\phi \qquad (3)
\end{aligned}$$

Eq. (3) does not contain λ. Hence, correct estimation of $Q(\Theta)$ does not depend on these elements.

If one had the query prior to the analysis, it would be possible to determine the partition of Θ into λ and ϕ, and derive an inference procedure that targets the integral in Eq. (3). In contrast, the proposed procedure assumes that the queries are posed after the training. Therefore we first learn Θ via $P_G(\Theta|x_o)$, and then sample Θ (both of its λ and ϕ components) during sampling-based inference on the mutilated graph. For example, suppose $Q(\Theta) = P_G(y|do(x'), \Theta)$. Then inference may take the form of a sampling procedure targeting the following integral:

$$\begin{aligned}
E(Q(\Theta)) &= \int_{\Theta} P_G(y|do(x'), \Theta) P_G(\Theta|x_o) d\Theta \\
&= \int_{\lambda,\phi} P_{G_{\bar{X}}}(y|x', \lambda, \phi) P_G(\lambda, \phi|x_o) d\lambda, d\phi
\end{aligned}$$

If $Q(\Theta)$ is identified, we know that λ does not impact the result, despite the fact that it is sampled in the inference procedure. When $Q(\Theta)$ is unidentified, then $Q(\phi, \lambda) \neq Q(\phi)$. The posterior is be spread around partial identification bounds if the identified part ϕ constrains λ in any way. □

Complexity of the algorithm Traditional approaches to causal inference construct the estimator after the query is specified. The proposed approach amortizes most of the computational work into a single training step inferring $P_G(\theta|\mathbf{x}_o)$, performed only once for all the parameters θ in the model to answer an arbitrary number of (identifiable) causal queries.

The training step (whether it is inferring $P(\theta|\mathbf{x}_o)$ or a point estimator of $\theta|\mathbf{x}_o$) depends on the inference algorithm. To evaluate its computational complexity, we assume the practical case of stochastic variational inference (SVI) where the proposal distribution is specified such that inference is exact. Inference with SVI takes advantage of state-of-the-art training methods, though in general it is not an exact technique. However, the trade-offs between exact and approximate inference are well known.

An exact alternative to SVI is the Hamiltonian Monte Carlo (HMC). The complexity of exact probabilistic inference on the graph depends on the parameterization of each factor $P(X|pa(X), \theta)$ in the joint distribution.

Causal LVMs with misspecified latent structure If a causal LVM has a misspecified latent structure with respect to the true LVM, but both latent projections result in the same ADMG, they will both produce the same estimand for an identifiable causal query. Choosing from the right set of LVM is a less restrictive constraint than choosing exactly the right LVM.

4 CASE STUDIES

4.1 Overall inference and evaluation strategy

For each case study, we used Causal Fusion to investigate the identifiability of the causal queries and derive do-calculus-based estimands. See supplementary material for do-calculus-based estimands for each case study. Inference was performed with HMC, and implemented in `Stan` [36] [2].

Let $\mu = E[\text{effect}|do(\text{cause})] - E[\text{effect}]$ and $\hat{\mu} = \hat{E}[\text{effect}|do(\text{cause})] - E[\text{effect}]$ be the true and the estimated average causal effects. For each case study, we specify the true values of model parameters and simulate observational data. Next, we simulate interventional data by fixing the value of the cause to its intervened value. μ is then calculated as the mean effect in the interventional data, and $\hat{\mu}$ is estimated by forward sampling on the mutilated graph. We evaluate the performance of the estimation in terms of the distribution of absolute error

[2] *https://github.com/srtaheri/LVMwithDoCalculus*

$AE = |\hat{\mu} - \mu|$ over 10 observational and 10 interventional datasets, and compare the box plots of them to those obtained with alternative estimators applicable to the specific causal query. To evaluate the robustness of the estimation to model misspecification, we also compare the performance of the proposed approach on a causal LVM trained with the true DAG structure (the true LVM) to a misspecified LVM, i.e. causal LVM trained with a misspecified latent structure but same ADMG.

4.2 *Case study 1: The Multi-cause model*

The system [7, 37] proposed a causal LVM in Fig. 1(a). Their work highlights the attractiveness of applying causal LVMs to multidimensional problems common in machine learning. E.g., variational autoencoder could capture the relationship between observed $\mathbf{X}$ and latent $\mathbf{U}$, where $P(\mathbf{x}|\mathbf{u})$ is learned by the decoder, and $P(\mathbf{u}|\mathbf{x})$ by the encoder. Their work also illustrates the challenge of causal identifiability in LVMs. The authors argued that it is possible to identify $P(y|do(\mathbf{x}))$ by relying on the multidimensionality of the latent variable $\mathbf{U}$. However, [38] showed that this is not true in general, except for cases with strong parametric assumptions. Case study 1 is similar to an example in [38]. It demonstrates that by extending the causal LVM in Fig. 1(a) with a mediator M as in Fig. 1(b), the causal effect of $\mathbf{X}$ on Y becomes non-parametrically identified by the do-calculus-based front-door criterion. As the result, it can be estimated correctly by an exact graph-based probabilistic inference on the mutilated graph.

Causal query of interest is $E[Y|do(\mathbf{X} = 0)]$.

Data are generated with three latent variables and five causes, where $\mathbf{U}$ follows Normal distribution and the remaining variables follow Bernoulli distribution with logit parameterization.

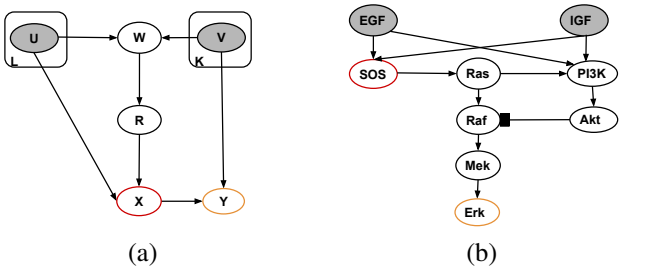
True and misspecified LVM The true LVM is as in Fig. 1(b) with 3 latent variables and 5 causes. The misspecified LVM only has one latent variable U. Non-informative $\mathcal{N}(0, 10)$ priors are used for all the parameters.

Accuracy of estimation of the causal effect is summarized in Fig. 3(a). Since the model has multiple causes, the proposed approach can only be compared to the plug-in estimator. Estimates based on the true LVM and the misspecified LVM outperform the plug-in estimator. Estimates based on the true LVM preforms best. Estimates based on the misspecified LVM converge to that of the true LVM as the number of data points increases.

4.3 *Case study 2 : The new Napkin problem*

The system Fig. 2(b) describes an observational study of patients with HIV, where treatment R affects CD4 cell counts X, and W is a known disease history, that affects the treatment. $\mathbf{U}$ and $\mathbf{V}$ are sets of latent confounders, such as underlying comorbidities, and Y is the disease outcome. This model requires a non-trivial application of the do-calculus, as we cannot block the back-door path, and the front-door criterion does not hold [39, 40, 10, 33].

Casual query of interest is the average outcome of disease after an intervention on CD4 cell counts, i.e. $E[Y|do(X = 0)]$. By intervening on X, the mutilated graph removes all the incoming edges to X. Hence the causal effect on Y only depends on X and V.



(a) (b)

Figure 2: Node representations are as in Fig. 1. (a) The Napkin model. X is the target of the intervention. Y is the effect. (b) The Signaling model. Nodes are proteins. Pointed/flat-headed edges are relationships of type *increase*/*decrease*. SOS is the target of intervention. Erk is the effect.

Data for the root nodes are simulated as Normal. W is simulated from a Gamma distribution. R and Y are simulated from a Normal distribution. X is simulated from a Bernoulli distribution with logit parametrization.

True and misspecified LVM The true LVM is modeled with the DAG in Fig. 2(b). The misspecified LVM has two latent causes of W and Y instead of one. Non-informative $\mathcal{N}(0, 10)$ priors are used for all the parameters.

Accuracy of estimation of the causal effect is summarized in Fig. 3(b). Even though this case study has one cause and one effect, the `ananke` library is unable to estimate the causal query with any of its estimators. Hence, we only compare the causal LVM to the plug-in estimator. The relative performance is as in case study 1.

4.4 *Case study 3: The Signaling model*

The system The insulin-like growth factor (IGF) signaling system in Fig. 2(b) regulates growth and energy metabolism of a cell. It is activated by external stimuli IGF and EGF. Nodes in the system represent kinase proteins, and edges represent the effect that the upstream kinase has on the downstream kinase's activity. IGF, EGF, and $PI3K$ are latent. This case study does not satisfy the back-door or the front-door criteria, and has a non-trivial data generation process defined by prior biological knowledge [41].

Causal query of interest is an ideal intervention fixing SOS to 70. We are interested in $E[Erk|do(SOS = 70)]$.

Data mimics the process of collecting observational and interventional data. Since dynamics of this system are well characterized in form of stochastic differential equations (SDE) [42], we generate observational data by simulating from an SDE. We set the initial amount of each protein molecule to 100, and generate subsequent observations via the Gillespie algorithm [43] in the *smfsb* [44] R package. Replicates are generated by randomly initializing EGF and IGF. Interventional data are generated similarly, while fixing $SOS = 70$.

True and misspecified LVM Since the latent $PI3K$ kinase has parents, we can avoid learning its parameters by transforming the network, removing $PI3K$, and re-directing all its incoming edges into Akt [45]. The true LVM is modeled with this transformed DAG. Probability distributions at each root

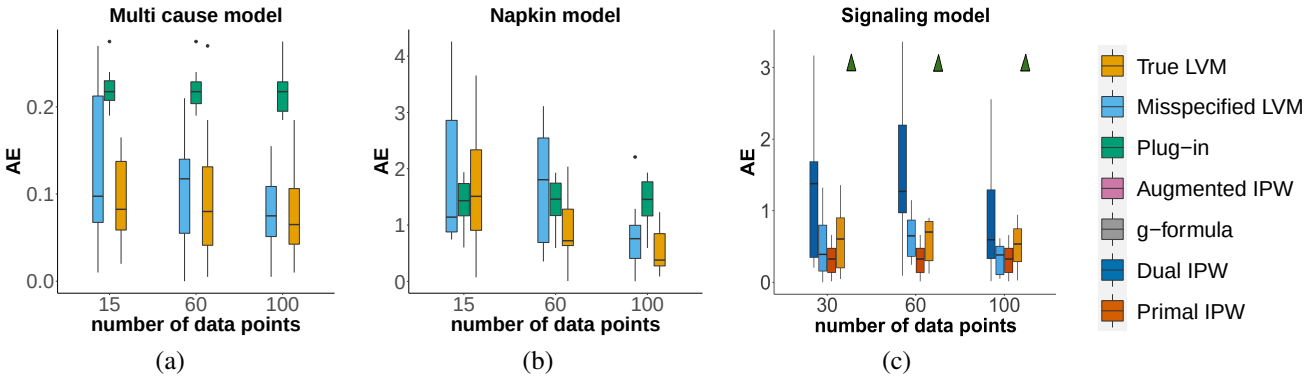


Figure 3: Absolute error of average causal effect estimation (AE). (a) The Multi-cause model. (b) The Napkin model. (c) The Signaling model. Triangles refer to plug-in estimates, for which AE exceeds 3 for all numbers of data points. In the case of the misspecified LVM with 100 data points, one dataset has AE=9, outside of the y axis limit.

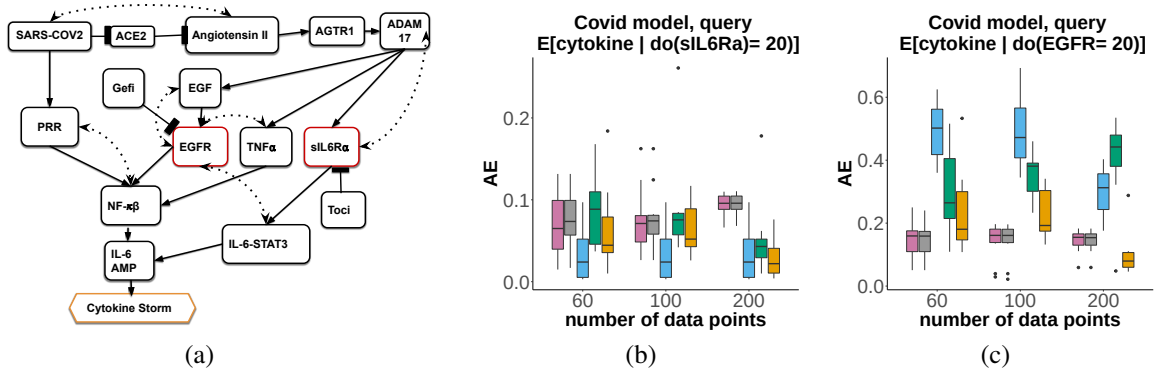


Figure 4: The Covid model. Boxplot colors are as in Fig. 3. (a) The causal LVM. Nodes are proteins, pointed/flat-headed edges are relationships of type *increase/decrease*, dotted edges indicate presence of latent variables. *sIL6Rα* and *EGFR* are targets of intervention. Cytokine Storm is the effect. (b) AE for $E[Cytokine|do(sIL6R\alpha = 20)]$. (c) AE for $E[Cytokine|do(EGFR = 20)]$.

node are $\mathcal{N}(\mu_r, \sigma_r)$. Probability distributions at each non-root node $\mathcal{N}(\frac{100}{1+exp(\theta' Pa(X)+\theta_0)}, \sigma_X)$ are motivated by common biological practice, where simple biomolecular reactions are modeled with Hill function [46] and approximated with a sigmoid. For a node X with q parents, $Pa(X)$ is a $q \times 1$ vector of measurements on the parent nodes, θ' is a $1 \times q$ vector of unknown parameters, and θ_0 is an unknown scalar bias parameter. Non-informative $\mathcal{N}(0, 10)$ priors are used for all the parameters, with the constraint that the parameter weight θ in the sigmoid is positive for the relationships of type increase and negative for relationships of type decrease. The misspecified LVM has a similar structure, while only including *EGF* and *PI3K* as latent and omitting *IGF*.

Accuracy of the estimator is summarized in Fig. 3(c). The do-calculus-based formula for the plug-in estimator, obtained using Causal Fusion platform, includes many conditional distributions. The estimator make parametric assumptions for each distribution and performs poorly. The `ananke` library suggests the Dual IPW and the Primal IPW as the best alternatives. While the Dual IPW performs poorly, the primal IPW performs best, slightly better than the true LVM and the misspecified LVM. The estimates by primal IPW, true LVM and misspecified LVM converge as the number of data points increases.

4.5 *Case study 4: The Covid model*

The system This small-scale expert system showcases the ability of a causal LVM to answer multiple causal queries after a single instance of training. It models activation of Cytokine Release Syndrome (cytokine storm or CytokineStorm), known to cause tissue damage in severely ill SARS-CoV-2 patients [47], Fig. 4(a). The simultaneous activation of the nuclear factor kappa-light-chain-enhancer of activated B cell (NF-κB or NF-κB) and Interleukin 6 STAT3 Complex (IL6-STAT3 or IL6-STAT3) initiates a positive feedback loop known as Interleukin 6 Amplifier (IL6-AMP or IL6-AMP), which in turn activates a cytokine storm [48].

The network was extracted from COVID-19 Open Research Dataset (CORD-19) [49] document corpus using the Integrated Dynamical Reasoner and Assembler (INDRA) [50] workflow [41], and by quering and expressing the corresponding causal statements in the Biological Expression Language (BEL) [51].

Presence of latent variables was determined by querying pairs of entities in the network for common causes in the corpus.

Causal queries examine the ability of two different drugs to prevent the cytokine storm. Tocilizumab (Toci or Toci) is an immunosuppressive drug that targets $sIL6R\alpha$ and blocks the IL6 signal transduction pathway [52]. Gefitinib (Gefi or Gefi) is an Epidermal Growth Factor Receptor (EGFR or EGFR) inhibitor, which blocks $EGFR$. The first causal query examines the effect of Toci by setting its target $sIL6R\alpha = 20$ (low value), and considering $E[Cytokine|do(sIL6R\alpha) = 20)]$. This causal query is identifiable using the backdoor criterion. The second causal query examines the effect of Gefi by setting its target $EGFR = 20$, and considering $E[Cytokine|do(EGFR = 20)]$. This causal query is not identifiable via either the backdoor or the front-door criterion, but can be identified via the do-calculus.

Data for the root nodes were simulated from a Normal distribution. Simulation of the non-root nodes was motivated by the same biological practice as in case study 3, and simulated by the same procedure. Cytokine storm has a Bernoulli distribution with logit parameterization.

True and misspecified LVM The true LVM contains two latent variables between $SARS - CoV - 2$ and $AngiotensinII$, $ADAM17$ and $sIL6R\alpha$, and PRR and $NF{-}\kappa B$, and one latent variable for each remaining dotted edge in Fig. 4(a). The misspecified LVM only has one latent variable for each dotted edge.

Accuracy of the estimators for the two queries is summarized in Fig. 4 (b)-(c). In addition to the plug-in estimator, the `ananke` library suggests the g-formula and the Augmented IPW as the best alternative estimators for each query. For the first query the misspecified LVM performs best. This may be due to the fact that it is less complex, and easier to train by HMC. Performance of the true LVM approaches that of the misspecified model as the number of observations increases. The plug-in, the g-formula and the Augmented IPW estimators perform slightly worse.

To estimate the second query, the models corresponding to the alternative estimators were retrained from scratch. In contrast, we do not retrain the LVM, but simply sample from a different mutilated version of the trained model. At the same time, the second query produces a different mutilated model and the associated estimand, and therefore evaluation on this query produces overall higher absolute errors and different relative performance. Unlike for the first query, the true LVM performs best. The g-formula and the augmented IPW perform the same or better than the plug-in estimator and the misspecified LVM.

5 DISCUSSION

This manuscript demonstrates the value of using a Bayesian approach on a trained causal LVM, to answer causal queries that are identifiable via the do-calculus rules. This is particularly useful in settings where multiple *ad hoc* queries arise after model training. The existing approaches for causal query estimation typically require rebuilding the underlying model from scratch. If the model incorporates deep learning architectures, this process must cope with nonlinear relationships between high-dimensional variables, and starting from scratch is computationally expensive. In contrast, the LVM, once trained, can answer an arbitrary number of queries, and is easily extendable to new variables in the model. The four case studies showed that the proposed approach is applicable to more situations than many alternative estimators, with satisfactory performance.

From Bayesian perspective, the estimator of the causal query can be thought of as a posterior predictive statistic. In practice, we need not be strictly Bayesian in training the parameters θ. The same approach can be implemented with alternative point estimate of model parameters, depending on bias, variance, and computational cost trade-offs.

The proposed approach has limitations. The distribution obtained by marginalizing out the latent variables may be intractable or contain singularities [53]. An LVM with misspecified latent structures may also entail different constraints on the joint marginal probability of the variables that were observed during training [21, 22]. As a result, models with different latent structure may differ in terms of fit to the training data or precision of their causal inference. These difficulties may be navigated with traditional model evaluation techniques, such as posterior predictive checks and model selection statistics. Exploring these issues is subject of our future work.

Acknowledgements We thank Carlos Cinelli and Alexander D'Amour for their useful comments and suggestions. Support for Jeremy Zucker was provided by the PNNL Laboratory-directed R&D Data-Model Convergence Initiative and the Mathematics for Artificial Reasoning Systems Initiative. PNNL is operated for the DOE by Battelle Memorial Institute under Contract DE-AC05-76RLO1830.

REFERENCES

[1] David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203, 2014.

[2] Diederik P Kingma and Max Welling. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, 2014.

[3] Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science ..., 1982.

[4] Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society B*, 73:123, 2011.

[5] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303, 2013.

[6] G Casella and RL Berger. Statistical inference, Duxbury, 2002.

[7] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*.

[8] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669, 1995.

[9] Judea Pearl. Bayesian analysis in expert systems: comment: graphical models, causality and intervention. *Statistical Science*, 8:266, 1993.

[10] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[11] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

[12] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74:981, 2007.

[13] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[14] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[15] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219, 2006.

[16] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv*, jan 2017.

[17] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of Uncertainty in Artificial Intelligence*, UAI'06, 2006.

[18] Santtu Tikka and Juha Karvanen. Identifying causal effects with the R package causaleffect. *arXiv:1806.07161*, 2018.

[19] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345, 2016.

[20] Robin J. Evans. Margins of discrete bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

[21] Judea Pearl. On the testability of causal models with latent and instrumental variables. *arXiv*, feb 2013.

[22] Robin J. Evans. Graphical methods for inequality constraints in marginalized DAGs. In *International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, sep 2012.

[23] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, page 6446, 2017.

[24] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. The deconfounded recommender: A causal inference approach to recommendation. *arXiv:1808.06581*, 2018.

[25] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41, 1983.

[26] Judea Pearl, F Bacchus, P Spirtes, C Glymour, and R Scheines. Probabilistic reasoning in intelligent systems: Networks of plausible inference. 1995.

[27] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2, 2006.

[28] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393, 1986.

[29] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[30] James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

[31] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

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

[33] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[34] Jaron JR Lee and Ilya Shpitser. Identification methods with arbitrary interventional distributions as inputs. *arXiv:2004.01157*, 2020.

[35] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. *arXiv:2004.04872*, 2020.

[36] Stan Development Team et al. RStan: the R Interface to Stan. R package version 2.17. 3, 2018.

[37] Rajesh Ranganath and Adler Perotte. Multiple causal inference with latent confounding. *arXiv:1805.08273*, 2018.

[38] Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv:1902.10286*, 2019.

[39] Jouni Helske, Santtu Tikka, and Juha Karvanen. Estimation of causal effects with small data under implicit functional constraints. *arXiv:2003.03187*, 2020.

[40] Michael D Hughes, Michael J Daniels, Margaret A Fischl, Soyeon Kim, and Robert T Schooley. Cd4 cell count as a surrogate endpoint in hiv clinical trials: A meta-analysis of studies of the aids clinical trials group. *Aids*, 12:1823, 1998.

[41] J. Zucker, K. Paneri, S. Mohammad-Taheri, S. Bhargava, P. Kolambkar, C. Bakker, J. Teuton, C. T. Hoyt, K. Oxford, R. Ness, and O. Vitek. Leveraging structured biological knowledge for counterfactual inference: A case study of viral pathogenesis. *IEEE Transactions on Big Data*, pages 1–1, 2021.

[42] F. Bianconi, E. Baldelli, V. Ludovini, L. Crino, A. Flacco, and P. Valigi. Computational model of EGFR and IGF1R pathways in lung cancer: a systems biology approach for translational oncology. *Biotechnology Advances*, 30:142, 2012.

[43] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81: 2340, 1977.

[44] D. Wilkinson. Package smfsb. 2018.

[45] Robin J Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43:625, 2016.

[46] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, 2019.

[47] Z. S. Ulhaq and G. V. Soraya. Interleukin-6 as a potential biomarker of COVID-19 progression. *Medecine et Maladies Infectieuses*, 50:382, 2020.

[48] T. Hirano and M. Murakami. COVID-19: A new virus, but a familiar receptor and cytokine release syndrome. *Immunity*, 52:731, 2020.

[49] L. Lu Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. CORD-19: The Covid-19 Open Research Dataset. *arXiv*, 2020.

[50] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13, 2017.

[51] T. Slater. Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today*, 19:193, 2014.

[52] C. Zhang, Z. Wu, J.-W. Li, H. Zhao, and G.-Q. Wang. Cytokine release syndrome in severe COVID-19: Interleukin-6 receptor antagonist Tocilizumab may be the key to reduce mortality. *International Journal of Antimicrobial Agents*, 55:105954, 2020.

[53] Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, Jan 2014.