# Causal Inference from an EM Learned Causal Model

Anna K Raichev(araichev@uci.edu), Jin Tian(jtian@iastate.edu),
Rina Dechter (dechter@ics.uci.edu)

Paper

## Overview

We propose an alternative paradigm for answering causal queries. The idea is to learn the full causal model from the observational data, and once a full model is available, the query can be answered by applying Probabilistic Graphical Models (PGM) algorithms. We show that when the diagram has a low induced-width this approach can be far more effective than the estimand-based approach.

Contributions:

1. A general scheme, Le4CI, for computing causal queries that utilizes well known algorithms for learning and inference, and the special case EM4CI that utilizes EM for learning.
2. An analysis of the scheme's theoretical properties, highlighting its challenges and benefits.
3. An empirical evaluation of EM4CI's performance on a set of synthetically generated benchmarks

## Problem

Given a causal diagram, a query $P(\boldsymbol{Y} \mid do(\boldsymbol{X}=x))$ and samples from the observed distribution, the task is to determine if the query can be answered (identifiability). If it is, then output the distribution of $P(\boldsymbol{Y} \mid do(\boldsymbol{X}=x))$.

**Current Practice:**

1. apply state of the art algorithms for identifiability. These are polynomial algorithms involving the graph and the query only. [Tian, 2002]
2. Generate an estimand, namely an algebraic expression for the query involving only probabilistic expressions over the visible variables.
3. Estimate the estimand from the observational data.

**Limitation**: functions in the estimand may be too large to estimate.

**Our approach:**

1. First learn a full causal model
2. Answer the query using PGM tools.

### Motivating Example

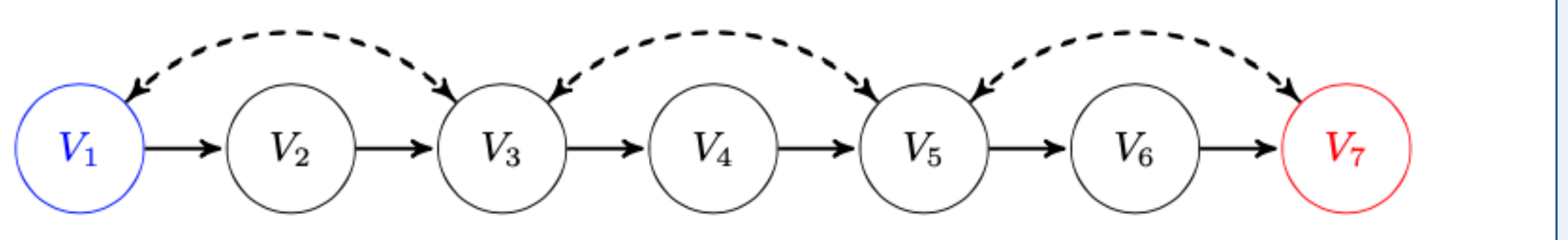


**Figure 1:** Chain Model with 7 observable variables and 3 latent variables

Using the estimated based approach we get the expression:

$$P(V_7 \mid do(V_1)) = \sum_{V_2,V_3,V_4,V_5,V_6} P(V_6 \mid V_1, V_2, V_3, V_4, V_5)P(V_4 \mid V_1, V_2, V_3)P(V_2 \mid V_1)$$

$$\times \sum_{V_1'} P(V_7 \mid V_1', V_2, V_3, V_4, V_5, V_6)P(V_5 \mid V_1', V_2, V_3, V_4)P(V_3 \mid V_1', V_2)P(V_1')$$

- Estimating this will take time exponential in number of variables
- However, the induced width of this model is only 2

## Background

**Structural Causal Model:** $\boldsymbol{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U})\rangle$

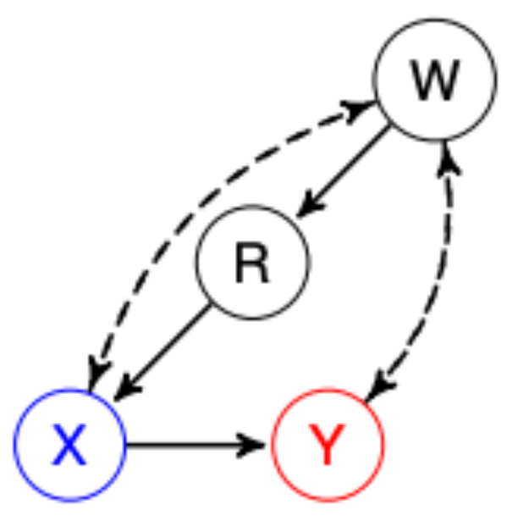
- $\boldsymbol{U} = \{U_1, \ldots, U_k\}$ set of unmeasurable latent variables
- $\boldsymbol{V} = \{V_1, V_2, \ldots, V_n\}$ set of observable variables
- $\boldsymbol{F} = \{f_i : V_i \in \boldsymbol{V}\}$ is a set of functional mechanisms $f_i$ that each determine the value $v_i$ of their corresponding $V_i$ as a function of $V_i$'s causal parents $PA_i \subseteq \boldsymbol{U} \cup \boldsymbol{V} \setminus V_i$
- $P(\boldsymbol{U})$ is a probability distribution over the exogenous variables

**Causal Diagram:** A SCM $\boldsymbol{M}$ can be associated with a directed graph $G = \langle \boldsymbol{V} \cup \boldsymbol{U}, E\rangle$ called a causal diagram. Each node in the graph uniquely corresponds to a variable in the SCM. There is an arc from node $X \in (\boldsymbol{U} \cup \boldsymbol{V})$ to node $V_i \in \boldsymbol{V}$ iff $X \in PA_i$

**Causal effect and the truncation formula**: We use $P(Y \mid do(X))$ to denote the distributions resulting from an intervention which fixes the value of $X$, and is called the causal effect of $do(X)$ on $Y$

$$P(\boldsymbol{V}, \boldsymbol{U} \mid do(\boldsymbol{X})) = \prod_{V_j \notin \boldsymbol{X}} P(V_j \mid PA_j) \cdot P(\boldsymbol{U})$$

Causal Diagram:



Blue variables are intervened on and red variables are the outcome variables corresponding to the query $P(Y \mid do(X))$

## Learning for Causal Inference

**Identifiability**

- Any two models that agree on the observational distribution and causal diagram will also agree on $P(\boldsymbol{Y} \mid do(\boldsymbol{X}=x))$

**Full Model for Causal Inference (FM4CI)**

```
Algorithm 1: FM4CI
input  : SCM $M = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U})\rangle$; causal query $Q = P(Y|do(X=x))$
output : $P(Y|do(X=x))$
  Step 1: $M_x \leftarrow \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F} \setminus \{f_i; V_i \in X\}, P(\boldsymbol{U})\rangle$
  Step 2: $result \leftarrow \text{inference}_{\text{PGM}}(M_X, Q)$
return $result$
```

1. truncate $\boldsymbol{M}$ into the causal model $\boldsymbol{M}_X$ by removing the function associated with $X$ and assigning $X=x$ in all functions where $X$ appears
2. apply a PGM algorithm to answer the associated query $P(\boldsymbol{Y} \mid \boldsymbol{X}=x)$

**EM for Causal Inference (EM4CI)**

```
Algorithm 2: EM4CI
input  : A causal diagram $\mathcal{G} = \langle \boldsymbol{U} \cup \boldsymbol{V}, E\rangle$; observable variables and their domains
         $\mathcal{D}_V = \{|\mathcal{D}(V_i)|\}$; assumed domain sizes $\mathcal{D}'_U = \{|\mathcal{D}'(U_i)|\}$; causal query
         $Q = P(Y \mid do(X=x))$; samples $S$ from the observational distribution $P(\boldsymbol{V})$
output : $P(Y \mid do(X=x))$
  Step 1: If $\neg$identifiable$(\mathcal{G}, Q)$, terminate.
  Step 2: $\mathcal{B} \leftarrow \text{EM}(\mathcal{G}, \mathcal{D}_V, \mathcal{D}'_U, S)$
  Step 3: $result \leftarrow \text{FM4CI}(\mathcal{B}, Q)$
return $result$
```

1. Check if query is identifiable
2. Using samples from the observed distribution $P(\boldsymbol{V})$ to learn a full causal Bayesian network $B$ consistent with $(\mathscr{G}, P(\boldsymbol{V}))$ using the EM algorithm
3. Use **FM4CI** to employ PGM inference techniques over the truncated learned model $B_{X=x}$ to compute $P(\boldsymbol{Y} \mid \boldsymbol{X}=x)$
4. Return $P(\boldsymbol{Y} \mid do(\boldsymbol{X}=x))$

**Complexity**

- Time and memory are exponential in the induced width

## Benefits & Challenges

**Challenges:**

1. In order to learn the full model we need to assume a domain size for the latent variables
2. There exists theoretical bounds on sufficient domain sizes. However the bounds are very conservative & can be very large to be practical [J. Zhang et al, 2022]
3. EM algorithm can be slow and converge to incorrect local optima in high dimensional space

**Benefits:**

1. Learning phase only needs to be performed once to answer *any* identifiable of form $P(\boldsymbol{Y} \mid do(\boldsymbol{X}=x))$; traditionally a new estimand would need to be derived for each query
2. Utilize the breadth of tools developed for graphical models
3. Expressions can be computationally intensive even for small induced width models, where learning is easy

## Experimental Setup

**Benchmarks**

- Each benchmark includes a causal diagram, a query, and observational data synthetically generated from the full model
- Used a range of domain sizes of the observed and latent variables

**Performance Measures**

- To evaluate the error of $P(\boldsymbol{Y} \mid do(\boldsymbol{X}=x))$, we use the mean absolute deviation (*mad*) and mean relative deviation (*mrd*)
- To evaluate the fitness of the learned model relative to the data we use the average log likelihood (*LL*)

## Notation

- Capital letters ($X$) represent variables, & small letters ($x$) represent their values. Boldfaced capital letters ($\mathbf{X}$) denote a collection of variables
- $n = |\mathbf{V}|$, $d = |D(V)|$, $k = |D(U)|$ in the true model, and $k_{hyp} = |D'(U)|$ the hypothesized domain of the learned model

## Empirical Analysis

### Baseline Comparison: Plug In Method

- Generates an estimand and the empirical conditional probabilities are computed from observational quantities
- Will converge to the exact result given enough samples

**(a)** 100 samples $d=2, k=10, k_{hyp}=16$

| Model | True Value | EM4CI (mad, time(s)) | Plug-in (error, time(s)) |
|---|---|---|---|
| 1 | 0.530 | (0.0374, 0.006) | (0.054, 0.108) |
| 2 | 0.481 | (0.0155, 0.025) | (0.0708, 0.029) |
| 3 | 0.519 | (0.022, 0.024) | (0.0393, 0.016) |
| 4 | 0.479 | (0.0186, 0.045) | (0.0459, 0.041) |
| 5 | 0.489 | (0.0085, 0.097) | (0.0123, 0.024) |
| 6 | 0.422 | (0.1097, 0.007) | (0.2915, 0.022) |
| 7 | 0.559 | (0.0192, 0.007) | (0.0804, 0.021) |
| 8 | 0.512 | (0.105, 0.007) | (0.0956, 0.027) |

### Competing Scheme: WERM

[Y. Jung et al., 2020]

- Learns causal effects by weighted empirical risk minimization
- State of the art method that focuses on estimating the quantities in the estimand using statistical methods

Comparison between WERM and EM4CI with $d=2$

| (Model, $|S|$) | True Value | WERM error | time(s) | EM4CI $k_{hyp}=2$ error | time |
|---|---|---|---|---|---|
| (1 ,1,000) | 0.0911 | 0.0071 | 18.7 | 0.002 | 0.043 |
| (8, 1,000) | 0.6972 | 0.1082 | 25.8 | 0.0769 | 0.069 |
| (9, 1,000) | 0.496 | 0.027 | 27.2 | 0.0305 | 0.054 |
| (1,10,000) | 0.0919 | 0.0031 | 32.6 | 0.0046 | 0.428 |
| (8, 10,000) | 0.699 | 0.11 | 47.7 | 0.0132 | 0.7 |
| (9, 10,000) | 0.491 | 0.001 | 44.1 | 0.0006 | 0.549 |

### Testing Different Hypothesized Domains

**(b)** Diamond Model: $n=65, d=2, k=2$
$P(V_{64}=0|do(V_0=0)) = 0.62424$

| $k_{hyp}$ | mad | mrd | LL | time(sec) |
|---|---|---|---|---|
| 2 | 0.012149 | 0.01946 | -32516.7 | 0.797 |
| 4 | 0.01467 | 0.0235 | -32524.5 | 0.897 |
| 8 | 0.008761 | 0.03296 | -32585.9 | 1.53 |
| 16 | 0.007541 | 0.05343 | -32886.4 | 12.94 |
| 32 | 0.051977 | 0.08326 | -33741.5 | 306.8 |
| 64 | 0.06889 | 0.11036 | -34721 | 7728.19 |

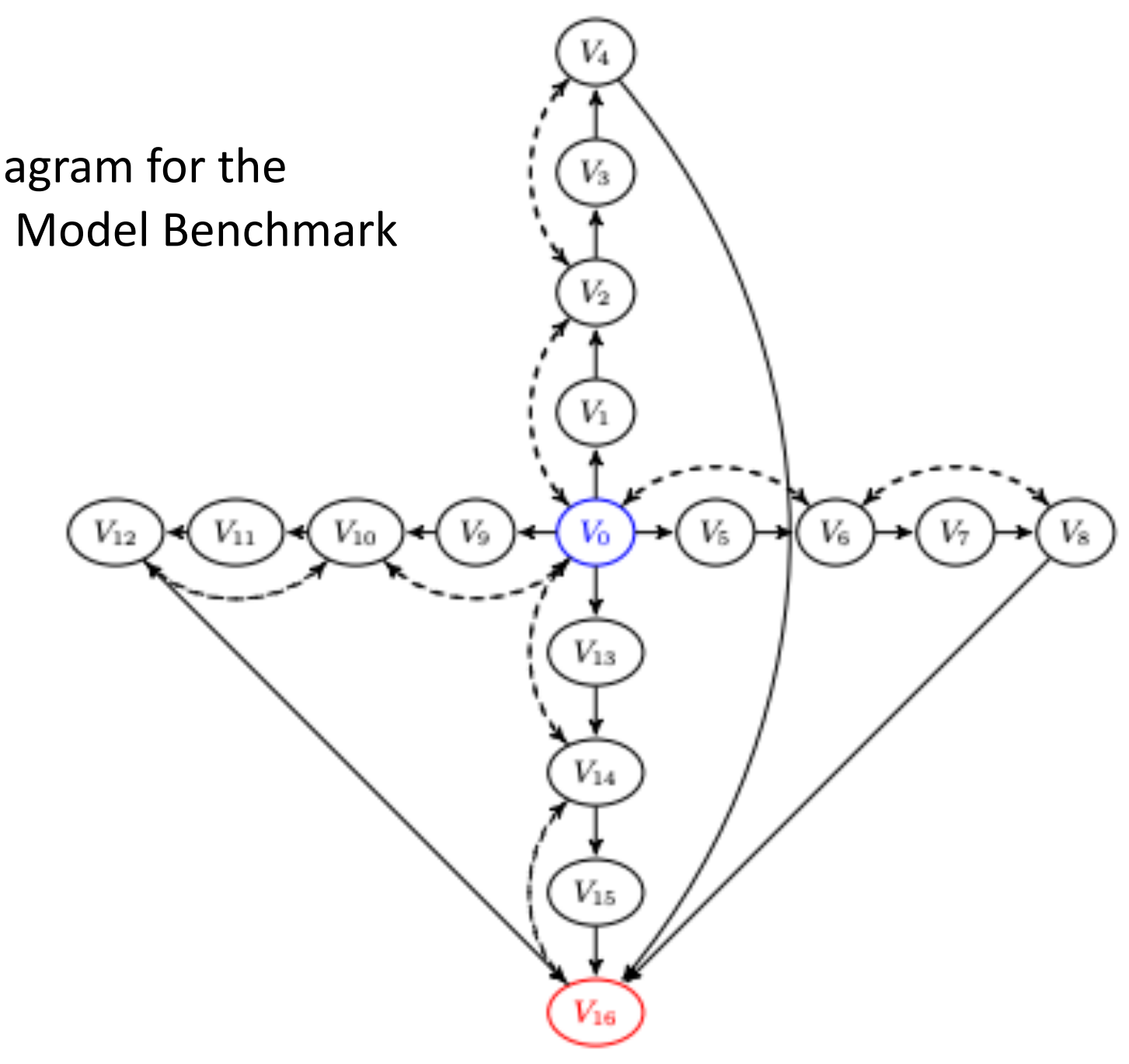
**(d)** Diamond Model: $n=65, d=2, k=8$
$P(V_{64}=0|do(V_0=0)) = 0.471171$

| $k_{hyp}$ | mad | mrd | LL | time(sec) |
|---|---|---|---|---|
| 2 | 0.00692 | 0.01469 | -37794.3 | 0.998 |
| 4 | 0.005744 | 0.01219 | -37794.7 | 0.878 |
| 8 | 0.003169 | 0.00673 | -37797.8 | 1.53 |
| 16 | 0.001969 | 0.00418 | -37815.6 | 12.85 |
| 32 | 0.009381 | 0.01991 | -37863.6 | 304.2 |
| 64 | 0.01578 | 0.03349 | -37916.5 | 7743.84 |

**(f)** Diamond Model: $n=65, d=2, k=16$
$P(V_{64}=0|do(V_0=0)) = 0.428615$

| $k_{hyp}$ | mad | mrd | LL | time(sec) |
|---|---|---|---|---|
| 2 | 0.009972 | 0.0233 | -37626.4 | 0.807 |
| 4 | 0.011396 | 0.02659 | -37627 | 0.887 |
| 8 | 0.013847 | 0.03231 | -37630.1 | 1.58 |
| 16 | 0.017812 | 0.04156 | -37638.6 | 14.351 |
| 32 | 0.023993 | 0.056 | -37657.6 | 311.6 |
| 32 | 0.023993 | 0.056 | -37657.6 | 311.6 |

Causal Diagram for the Diamond Model Benchmark



## Conclusion

- EM4CI was fast compared to other methods
- *Mad* and *mrd* were small on most benchmarks
- EM4CI is another tool for causal inference, not meant to replace the estimand based approach but used as an alternative when beneficial (low induced width models)

**Open Question**

- What is the best hypothesized domain to use?