Probabilistic Programming for constraint-based causal learning

Leonardo Casarsa and Vikash K. Mansinghka

Abstract

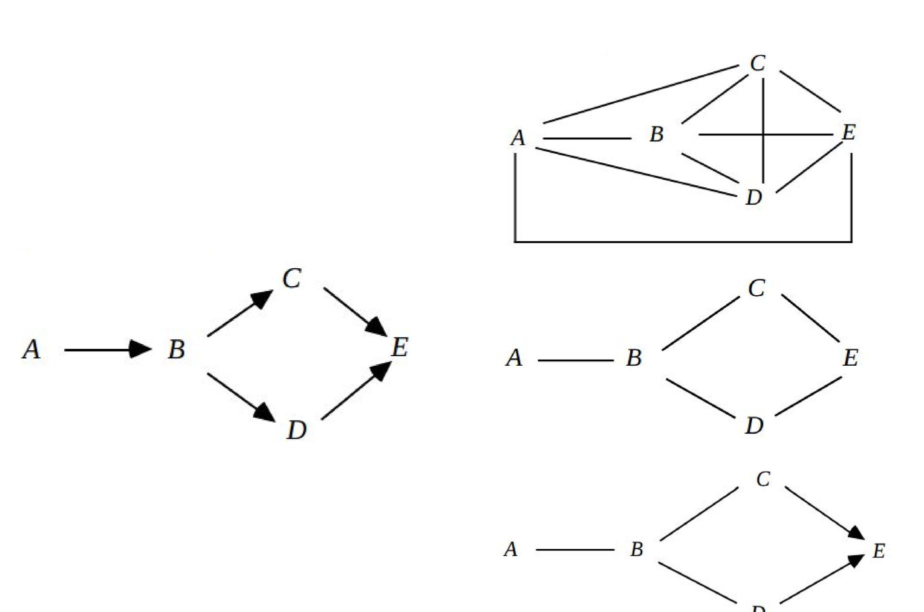
How can probabilistic programming facilitate causal inference? This poster presents one possible approach based on a computational formalization of conditional (in)dependence that can be assessed for black-box probabilistic programs whose source is at least partially unavailable. The BayesDB MML language is used to produce a posterior distribution over probabilistic programs conditioned on the data in an SQL database table, and causal extensions to BQL are introduced for querying the probable causal relationships that are compatible with the probability distribution that was inferred. This approach has potentially appealing features that we hope to explore in future work. Examples include (i) the ability to predict the likely effect of interventions when model structure is at least partially unknown (ii) formulations of causality and intervention that layer atop probabilistic programs, analogously to how the do-calculus layers atop of directed graphical models.

The PC Algorithm

Algorithm 1 Find Skeleton and Separation Set

Require: data table D with n columns, conditional independence test

1: $\mathcal{G} \leftarrow$ complete_graph(D) ▷ complete undirected graph on columns of D
2: $S \leftarrow \lambda\,(x, y) : \emptyset$ ▷ Initialize separation set
3: **for** $j \leftarrow 0$ to n **do** ▷ $|Z| = j$
4: **for** $(x, y) \in \{(x, y) : x \in \mathcal{G}, y \in adj(\mathcal{G}, x), |adj(\mathcal{G}, x) \setminus \{y\}| \geq j\}$ **do**
5: **for** $Z \in \{Z : Z \subset adj(\mathcal{G}, x) \setminus \{y\}, |Z| = j\}$ **do**
6: **if** $x \perp\!\!\!\perp y | Z$ **then**
7: $\mathcal{G} \leftarrow \mathcal{G} \setminus \{(x, y), (y, x)\}$ ▷ Remove edge between x and y from $\mathcal{G}$
8: $S(x, y) = S(y, x) \leftarrow Z$ ▷ Update separating set of x, y
9: **end if**
10: **end for**
11: **end for**
12: **end for**
% Intermediate state: $\mathcal{G}$ = skeleton, S = Separation Set
13: **for** $(x, w, y) \in \{(x, w, y) : w \in \mathcal{G}, x, y \in adj(\mathcal{G}, w), x \notin adj(\mathcal{G}, y), w \notin S(x, y)\}$ **do**
14: $\mathcal{G} \leftarrow \mathcal{G} \setminus \{(w, x), (w, y)\}$ ▷ Orient v-structure
15: **end for**
16: **repeat**
17: Orient, if possible, non v-structures
18: Orient edges to avoid cycles
19: **until** $\mathcal{G}$ stops changing
20: **return** $\mathcal{G}$ ▷ Causal CPDAG



Altered from [1]. Left: Graph of example data generating program. **Right top**: Complete undirected graph. **Right medium**: Skeleton. **Right bottom:** output of PC algorithm.

Conditional independence test for black-box probabilistic programs

A conditional independence test for black-box probabilistic programs can be computationally formalized with BayesDB. The BayesDB MML language is used to produce a posterior distribution over probabilistic programs conditioned on the data in an SQL database table, each sample of the distribution being a Generative Population Model, that is, a probabilistic object that provides a *sampler* of a random variable from its conditional distribution and an *assessor* of its conditional density.

ESTIMATE CMI OF X WITH Y GIVEN Z USING MODEL 1

$$\{\hat{z}_i\}_{i=1}^{l} \sim \{Z\}$$
$$\{\hat{x}_{ij}, \hat{y}_{ij}\}_{j=1}^{n} \sim \{X, Y\}|\{Z = \hat{z}_i\}, \text{ for } i = 1, \ldots, l$$
$$C\hat{M}I(X, Y|Z) \approx \frac{1}{ln} \sum_{i,j} \log \frac{p(X = \hat{x}_{ij}, Y = \hat{y}_{ij}|Z = \hat{z}_i)}{p(X = \hat{x}_{ij}|Z = \hat{z}_i)p(Y = \hat{y}_{ij}|Z = \hat{z}_i)}$$

SIMULATE (CMI OF X WITH Y GIVEN Z)

$$\{s_k\}_{k=1}^{m} \sim \{CMI(X; Y|Z)\}$$
$$\hat{p}(CMI(X; Y|Z) < \epsilon) = \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}_{s_k < \epsilon}$$

$$X \underset{\beta,\epsilon}{\perp\!\!\!\perp} Y|Z \iff \hat{p}(CMI(X; Y|Z) < \epsilon) > \beta$$

Illustration

$$\{\hat{x}_i\}_{i=1}^{1000} \sim \mathcal{N}(0, 1)$$
$$\{\hat{y}_i\}_{i=1}^{1000} \sim \mathcal{N}(0, 1)$$
$$\{\hat{z}_i\}_{i=1}^{1000} \sim \hat{x}_i + \hat{y}_i + \mathcal{N}(0, 1)$$

```
INITIALIZE 64 MODELS FOR experiment_gpmcc
ANALYZE experiment_gpmcc FOR 100 ITERATIONS
SIMULATE (CMI OF X WITH Y)
SIMULATE (CMI OF X WITH Y GIVEN Z)
```

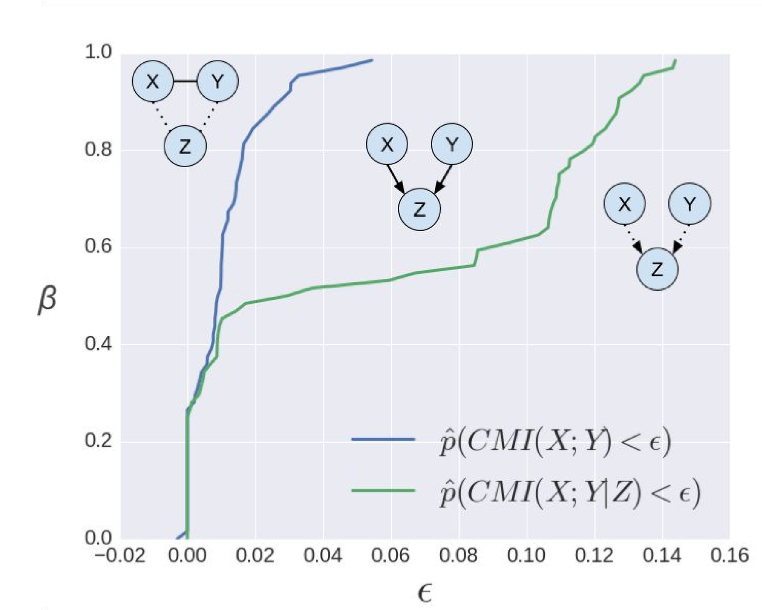


Fig 1: Preliminary results of CMI independence test on structure learning of Gaussian V-structure

$$\{\hat{x}_i\}_{i=1}^{1000} \sim \mathcal{N}(0, 1)$$
$$\{\hat{y}_i\}_{i=1}^{1000} \sim \mathcal{N}(0, 1)$$
$$\{\hat{z}_i\}_{i=1}^{1000} \sim \hat{x}_i + \hat{y}_i + \mathcal{N}(0, 1)$$

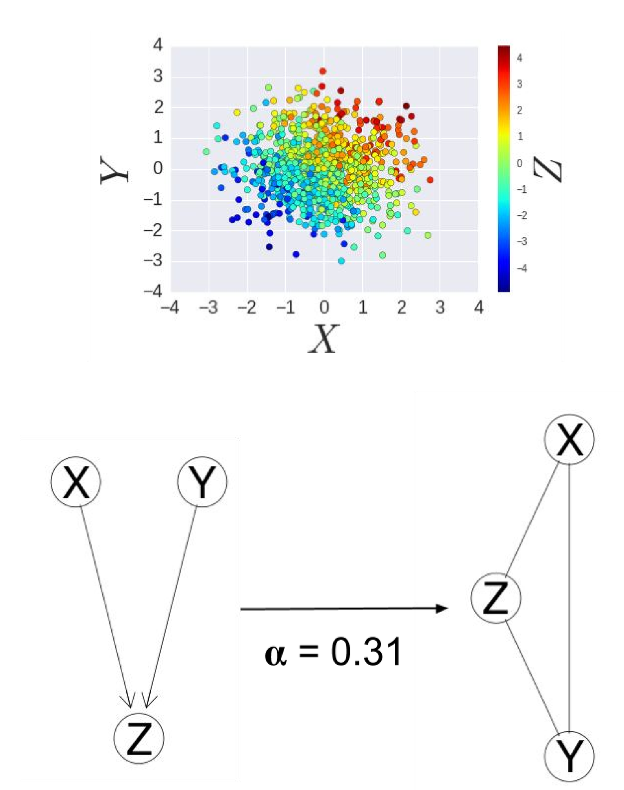


Fig 2: PC[1,2] using partial correlation on Gaussian V-structure

$$\{\hat{x}_i\}_{i=1}^{300} \sim \text{Bernoulli}(0.7)$$
$$\{\hat{y}_i\}_{i=1}^{300} \sim \text{Bernoulli}(0.4)$$
$$\{\hat{z}_i\}_{i=1}^{300} \sim \mathcal{N}(\mu_{\hat{x}_i, \hat{y}_i}, 1.1)$$

$$\text{where } \mu_{x,y} = \begin{cases} 0, & \text{if } x = 0, y = 0. \\ 4, & \text{if } x = 1, y = 0. \\ -4, & \text{if } x = 0, y = 1. \\ 8, & \text{if } x = 1, y = 1. \end{cases}$$

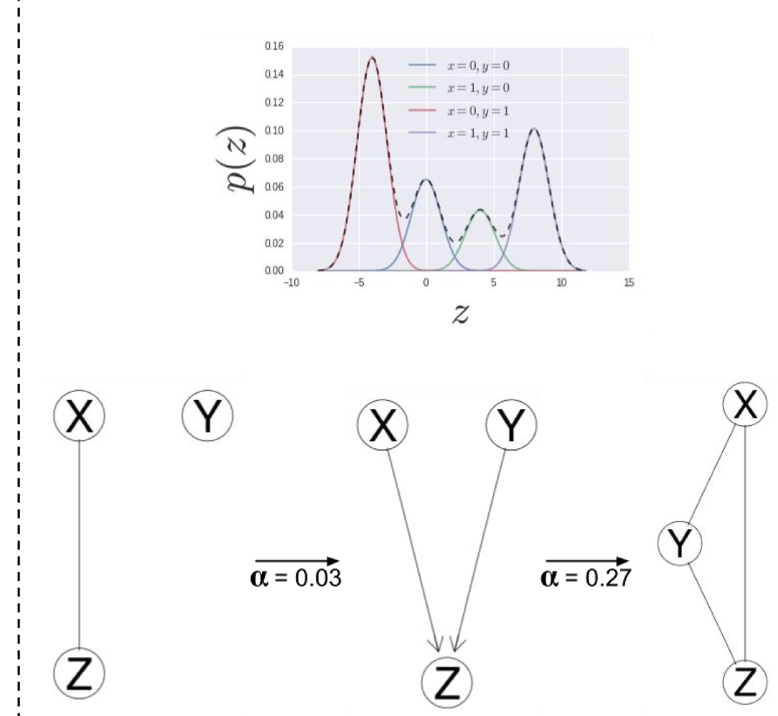
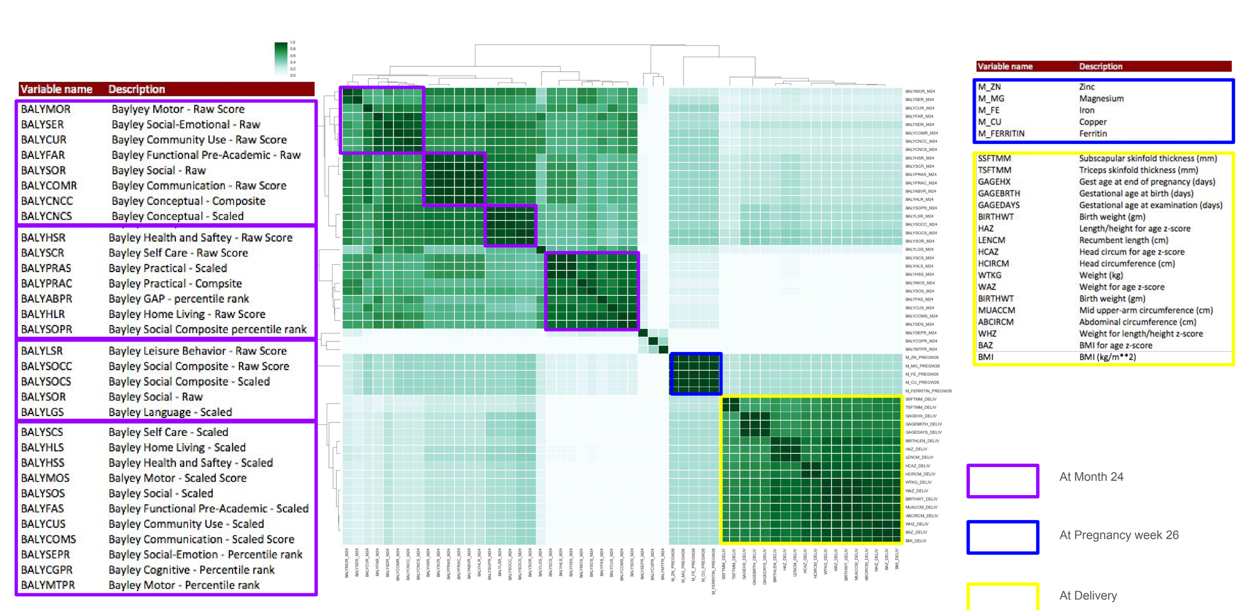
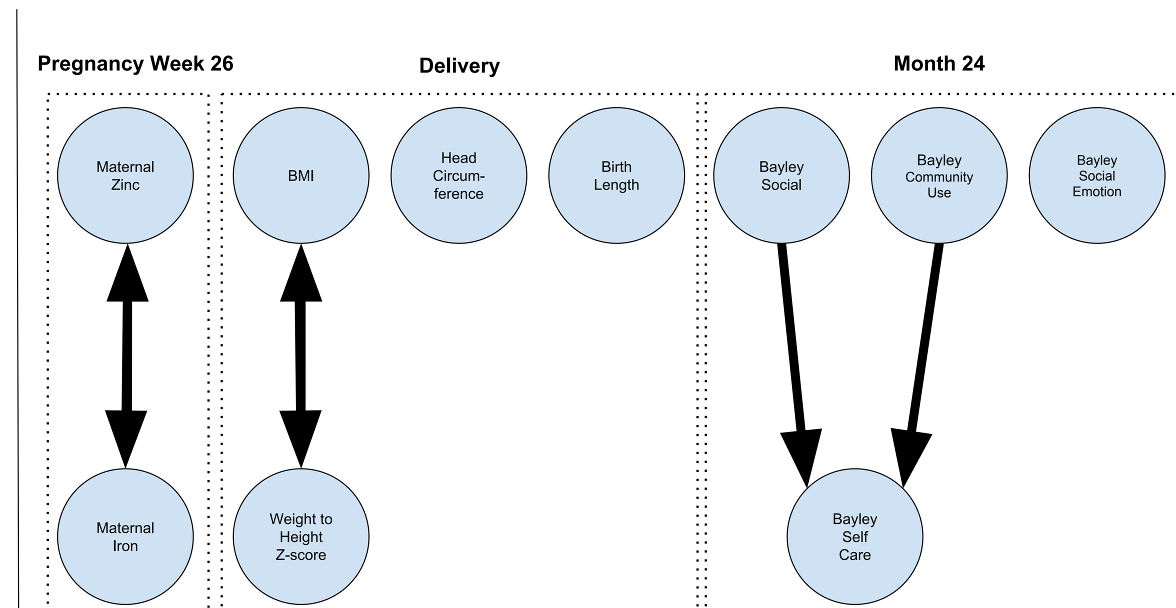


Fig 3: PC[1,2] using partial correlation on mixed types V-structure

Causal Network from Real Data



Heatmap of pairwise dependence probability from a subset of variables from the GUSTO database.



Output of PC algorithm[1,2] using partial correlation test with $\alpha = 0.01$ on a reduced subset of 10 variables from the GUSTO data base.

Goal: reading out causal summaries and/or posterior causal inferences of GPMs (in BQL)

```
CREATE POPULATION satellites FROM ucs_satellites.csv
```

```
CREATE GENERATIVE POPULATION MODEL sat_composite FOR satellites USING compositor(
  STATISTICAL TYPES GUESS(*) IGNORE name;

  OVERRIDING purpose GIVEN power_watts, class_of_orbit, country_of_operator, inclination_radians
     USING SCIKIT-LEARN CONDITIONAL GPM random_forest;

  OVERRIDING period_minutes GIVEN apogee_km, perigee_km AND EXPOSING cluster_id, noise
     USING VENTURESCRIPT CONDITIONAL GPM (
      (assume kepler (lambda (apogee perigee)
          (let ((GM 398600.4418) (earth_radius 6378)
                (a (+ (* .5 (+ (abs apogee) (abs perigee))) earth_radius)))
            (/ (* (* 2 3.1415) (sqrt (/ (pow a 3) GM))) 60))))

      (assume alpha (gamma 1 1))
      (assume cluster_sampler (make_crp alpha))
      (assume simulate_cluster_id (mem (r) (cluster_sampler)))

      (assume noise_sampler (mem (lambda (cluster) (make_nig_normal 1 10 1 20))))
      (assume simulate_noise (mem (lambda (r) ((noise_sampler (simulate_cluster_id r))))))

      (assume simulate_period_minutes (mem (lambda (r)
          (+ (kepler (get_input 'apogee r) (get_input 'perigee r)) (simulate_noise r))))));
    );

  OVERRIDING apogee_km, perigee_km AS CAUSAL ANCESTORS WITH period_minutes
  OVERRIDING perigee_km AS NOT CAUSAL ANCESTOR WITH apogee_km
  OVERRIDING apogee_km AS NOT CAUSAL ANCESTOR WITH perigee_km

INITIALIZE 16 MODELS FOR sat_composite;
ANALYZE sat_composite FOR 4 MINUTES;

ESTIMATE COLUMNS WHERE PROBABILITY OF CAUSAL ANCESTOR WITH perigee_km = 0.6

ESTIMATE CAUSAL SUMMARY OF sat_composite VIA CMI-PC WITH beta=0.95, epsilon=0.01

ESTIMATE CAUSAL SUMMARY OF sat_composite VIA CHISQ-PC WITH alpha=0.01
```

Acknowledgements and References

The authors would like to thank Feras Saad, Ulrich Schaechtle, Gregory Marton and Taylor Campbell for helpful feedback and discussions. This research was supported by DARPA (PPAML program, contract number FA8750-14-2-0004), IARPA (under research contract 2015-15061000003), the Office of Naval Research (under research contract N000141310333), the Army Research Office (under agreement number W911NF-13-1-0212), and gifts from Analog Devices and Google.

[1] P. Spirtes, C. Glymour & R. Scheines (2000). Causation, Prediction, and Search. *MIT Press*, Adaptive Computation and Machine Learning, second edition.

[2] M. Kalisch, M. Mächler, D. Colombo, M.H. Maathuis & P. Bühlmann (2012). Causal Inference Using Graphical Models with the R package pcalg. Journal of Statistical Software, 47(11), 1-26.

[3] V. Mansinghka, R. Tibbetts, J. Baxter, P. Shafto & B. Eaves (2015). BayesDB: A probabilistic programming system for querying the probable implications of data. arXiv preprint arXiv 1512.05006.