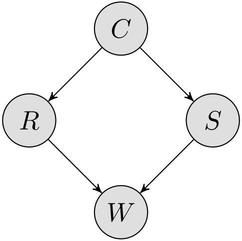
Learning Functional Causal Models with Generative Neural Networks

Presented by:
Haluk Dogan
https://haluk.github.io/
hdogan@vivaldi.net

Department of Computer Science
University of Nebraska-Lincoln

September 25, 2019

Introduction



C: Cloudy, R: Rainy, S: Sprinkler, W: WetGrass

$$
\begin{aligned}
R, E_S, E_W &\leftarrow U(0,1) \\
S &\leftarrow 0.5R + E_S \\
W &\leftarrow S + E_W
\end{aligned}
$$

Regression Solution

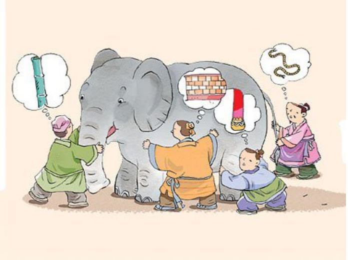
$$\hat{S} = 0.25R + 0.5W$$

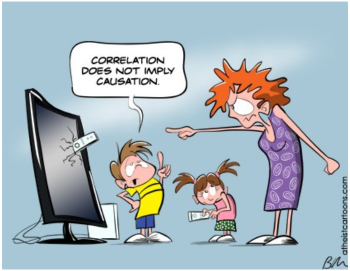
Derivation

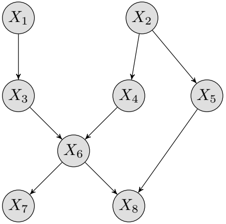
$S = B_1 R + B_2 W$
$S = 0.5R + E_S = W - E_W$
$E_S \sim E_W$ so $S \sim W$
$S = 0.5 \times U(0,1) + U(0,1)$

$$
\begin{aligned}
E(S \mid R, W) &= 0.5 \times E_R \times R + E_W \times W \\
&= 0.25R + 0.5W
\end{aligned}
$$

Introduction (cont'd)







$$
\begin{aligned}
P(X_1, \ldots, X_8) = {} & P(X_1)P(X_2)P(X_3 \mid X_1) \\
& P(X_4 \mid X_2)P(X_5 \mid X_2)P(X_6 \mid X_3, X_4) \\
& P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)
\end{aligned}
$$

Assuming variables are boolean:

Representation cost:

- Joint probability: 2^8
- BN: $2+2+4+4+4+8+4+8=36$

Introduction (cont'd)

Marginal Independence

$X \perp\!\!\!\perp Y$ iff

$$P(X,Y) = P(X)P(Y)$$
$$P(X \mid Y) = P(X)$$
$$P(Y \mid X) = P(Y)$$

Conditional Independence

$P(X \perp\!\!\!\perp Y \mid Z)$ iff

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$
$$P(X \mid Y,Z) = P(X \mid Z)$$
$$P(Y \mid X,Z) = P(Y \mid Z)$$

Introduction (cont'd)

Discovering causal relations requires performing experiments

- unethical
- expensive
- difficult to repeat

Interventation

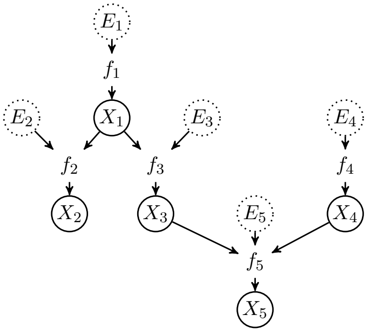
$\boldsymbol{X} = [X_1, \ldots, X_d]$
$do(X_i = x)$

Direct Cause

$X_i \to X_j$ iff
$P(X_j \mid do(X_i = x, \boldsymbol{X}_{\setminus ij} = \boldsymbol{c})) \neq P(X_j \mid do(X_i = x', \boldsymbol{X}_{\setminus ij} = \boldsymbol{c}))$

Outline

Functional Causal Model (FCM)

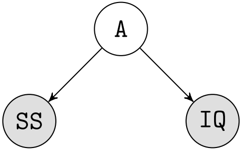


$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(X_1, E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

- FCM on a random variable vector $\boldsymbol{X} = [X_1, \ldots, X_d]$ is a triplet $(\mathscr{G}, f, \mathscr{E})$
- $X_i \leftarrow f_i(X_{\mathsf{Pa}(i;\mathscr{G})}, E_i)$, $E_i \sim \mathscr{E}$, for $i = 1, \ldots, d$
- E_i is used to account all unobserved variables and noise

Assumptions and Properties

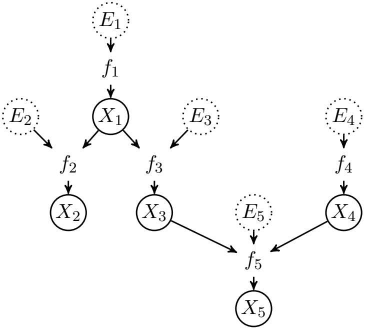
- Causal sufficiency assumption (**CSA**): common causes of all variables are measured



- $\mathtt{SS} \perp\!\!\!\perp \mathtt{IQ} \mid \mathtt{A}$
- Suppose `A` is unmeasured
- Data will only include independence statements not conditioned on `A`

Assumptions and Properties (cont'd)

- Causal Markov assumption (**CMA**):
  r.v $\perp\!\!\!\perp$ non-descendants (non-effects) | parents (direct causes)
  by Spirtes 2001
- For an FCM, this assumption holds if the graph is a DAG and error terms E_i in the FCM are independent on each other



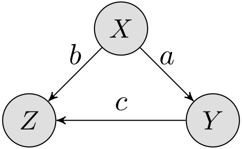
Assumptions and Properties (cont'd)

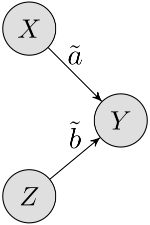
- Conditional independence relations in an FCM:
  If **CMA** applies, the data generated by the FCM satisfy all CI relations in $\boldsymbol{X}$ via d-separation by Pearl 2009

Faithfulness Assumption

There may be more CIs in data than present in the model

Assumptions and Properties (cond't)



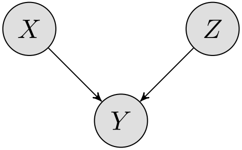


$$a \times c + b = 0, E. \sim N(0, \sigma_.^2)$$
$$X = E_x$$
$$Y = aX + E_Y$$
$$Z = bX + cY + E_Z$$
$$Z = -acX + c(aX + E_Y) + E_Z$$
$$Z = cE_Y + E_Z$$

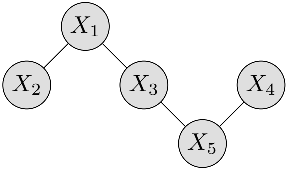
$$\tilde{a} = a, \tilde{b} = (b\sigma_Y^2)/(b^2\sigma_Y^2 + \sigma_Z^2), E. \sim N(0, \tau_.^2)$$
$$\tau_X^2 = \sigma_X^2$$
$$\tau_Y^2 = \sigma_Y^2 - (b^2\sigma_Y^4)/(b^2\sigma_Y^2 + \sigma_Z^2)$$
$$\tau_Z^2 = b^2\sigma_Y^2 + \sigma_Z^2$$
$$\Sigma = \begin{pmatrix} \sigma_X^2 & a\sigma_X^2 & 0 \\ a\sigma_X^2 & a^2\sigma_X^2 + \sigma_Y^2 & b\sigma_Y^2 \\ 0 & b\sigma_Y^2 & b^2\sigma_Y^2 + \sigma_Z^2 \end{pmatrix}$$

Assumptions and Properties (cont'd)

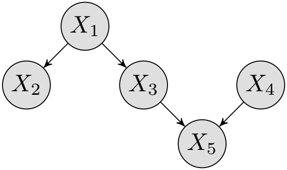
- v-structure property: $(X \not\perp\!\!\!\perp Z \mid Y)$



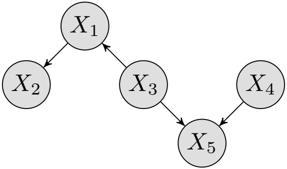
Learning the CPDAG



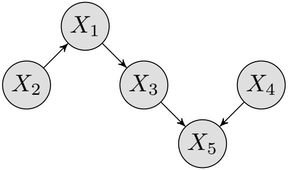
(a) Skeleton of a DAG



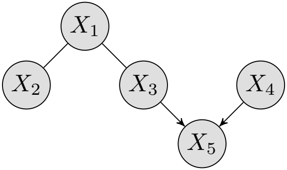
(b) Exact DAG



(c) Markov equivalent DAG



(d) Markov equivalent DAG

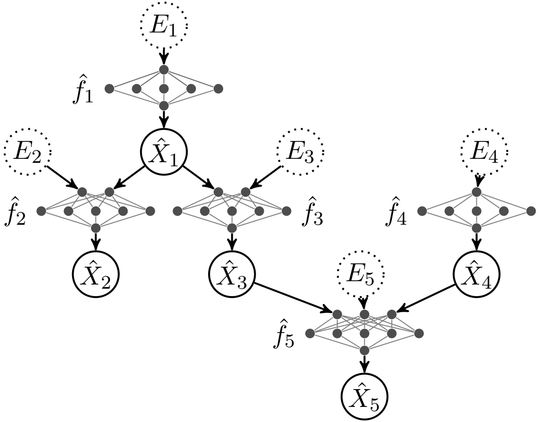


(e) CPDAG

Learning Algorithms

- **Constraint based**: Recover graph structure using tests of conditional independence
- **Score based**: Explore space of graphs while maximizing some scoring function defined relative to data
- **Hybrid**: Combination of constraint / score based methods
- **Pairwise**:
  - Restricting the class of functions allowed for causal mechanisms f_i and assuming a functional form
  - Regularize functions f_i with respect to local score and (empirically) helps the problem of identifiability

FCGNN



$$\begin{cases} \hat{X}_1 = \hat{f}_1(E_1) \\ \hat{X}_2 = \hat{f}_2(\hat{X}_1, E_2) \\ \hat{X}_3 = \hat{f}_3(\hat{X}_1, E_3) \\ \hat{X}_4 = \hat{f}_4(E_4) \\ \hat{X}_5 = \hat{f}_5(\hat{X}_3, \hat{X}_4, E_5) \end{cases}$$

$$\hat{X}_i = \hat{f}_i(\hat{X}_{\mathsf{Pa}(i;\mathscr{G})}, E_i) = \sum_{k=1}^{n_h} \bar{w}_k^i \sigma \left(\sum_{j \in \mathsf{Pa}(i;\mathscr{G})} \hat{w}_{jk}^i \hat{X}_j + w_k^i E_i + b_k^i \right) + \bar{b}^i$$

FCGNN (cont'd)

$$S(\mathscr{C}_{\hat{\mathscr{G}},\hat{f}}, \mathscr{D}) = \widehat{\mathsf{MMD}}_k(\mathscr{D}, \widehat{\mathscr{D}}) + \lambda|\widehat{\mathscr{G}}|$$

$$\widehat{\mathsf{MMD}}_k(\mathscr{D}, \widehat{\mathscr{D}}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\hat{x}_i, \hat{x}_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(x_i, \hat{x}_j)$$

- k: Gaussian kernel, $k(x, x') = \exp(-\gamma\|x - x'\|_2^2)$, differentiable
- $\lambda|\widehat{\mathscr{G}}|$ is a penalty term used for fair comparisons

Weight Optimization

$$\min_{\hat{\mathscr{G}}_i} \widehat{\mathsf{MMD}}_k(\mathscr{D}, (\hat{\mathscr{G}}_i, \hat{f}_1 \dots \hat{f}_d, \mathscr{E}))$$

- Adam optimizer
- Noise samples are drawn in each epoch (training and testing)

Structure Optimization

- Number of DAGs are super-exponential in size $|V|$ (Robinson 1977)
  - Structure optimization intractable
- Initial graph skeleton recovered by other methods such as feature selection (Yamada, Jitkrittum, Sigal, Xing, and Sugiyama 2014)
  - optimizing the edge orientations
  - Compare $S(\mathscr{C}_{X_i \to X_j, \hat{f}}, \mathscr{D}_{ij})$ and $S(\mathscr{C}_{X_j \to X_i, \hat{f}}, \mathscr{D}_{ij})$
  - keep the one gives the smaller score
  - Complexity $O(|E|)$
- Remove cycles from an initial graph
- Hill-Climbing (local search minimization)

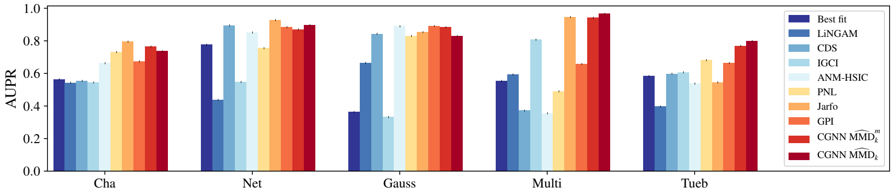
$$S_{X_i \to X_j} = S(\mathscr{C}_{\hat{\mathcal{G}} - \{X_i \to X_j\}, \hat{f}}, \mathscr{D}) - S(\mathscr{C}_{\hat{\mathcal{G}}, \hat{f}}, \mathscr{D})$$

Bivariate and Multivariate Causal Structures

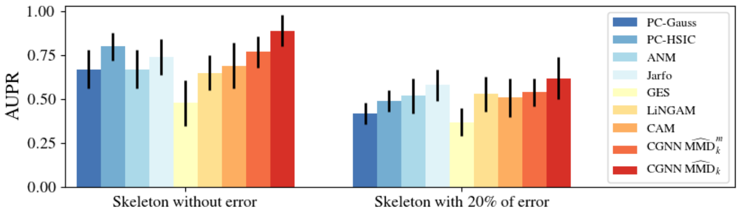
- Bivariate: 1500 samples
- **CE-Cha**: 300 continuous variable pairs
- **CE-Net**: 300 pairs generated with NN (random cause, e.g., exponential, gamma, lognormal, laplace)
- **CE-Gauss**: 300 pairs $Y = f_Y(X, E_Y), X = f_X(E_X)$
- **CE-Multi**: 300 artificial pairs (linear and polynomial)
    - post additive noise: $Y = f(X) + E$
    - post multiplicative noise: $Y = f(X) \times E$
    - pre-additive noise: $Y = f(X + E)$
    - pre-multiplicative noise $Y = f(X \times E)$
- **CE-Tueb**: 99 real-world cause-effect pairs
- Multivariate: 500 samples, 20 variables

Bivariate and Multivariate Causal Structures (cont'd)

Bivariate:



Multivariate:



Identifying v-Structures

500 samples

(a) skeleton (b) chain (c) reversed (d) v-structure

Score	non V-structures		V structure
	Chain str.	Reversed-V str.	V-structure
C_{ABC}	*0.122 (0.009)*	*0.124 (0.007)*	0.172 (0.005)
C_{CBA}	*0.121 (0.006)*	*0.127 (0.008)*	0.171 (0.004)
$C_{reversedV}$	*0.122 (0.007)*	*0.125 (0.006)*	0.172 (0.004)
$C_{V\,structure}$	0.202 (0.004)	0.180 (0.005)	**0.127** (0.005)

Questions

Questions?

References I

Pearl, J. (2009). *Causality*. Cambridge University Press. DOI: 10.1017/cbo9780511803161 (cit. on p. 10).

Robinson, R. W. (1977). “Counting unlabeled acyclic digraphs”. In: *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, pp. 28–43. DOI: 10.1007/bfb0069178 (cit. on p. 18).

Spirtes, P. (Mar. 11, 2001). *Causation, Prediction and Search*. MIT University Press Group Ltd. ISBN: 0262194406. URL: https://www.amazon.com/Causation-Prediction-Adaptive-Computation-Learning/dp/0262194406 (cit. on p. 9).

Yamada, M., W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama (Jan. 2014). “High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso”. In: *Neural Computation* 26.1, pp. 185–207. DOI: 10.1162/neco_a_00537 (cit. on p. 18).