**3.3 Causal Decision Making**

Dr Cheng Zhang (Microsoft Research)
https://www.microsoft.com/en-us/research/people/chezha/
http://www.dagitty.net/

**Casual machine learning**

- Key of decision making: “what if?”
- Understand the casual relationship: Causal Discovery
- Understanding the impact of the actions: Casual Inference

**If there is a (Partial) Causal Graph?**

- Do-calculus (Pearl)
- Potential outcome (Rubin)

**Causal discovery**

- Tasks: relationship between variables
- Directed Acyclic Graph
    - G = { V, E}
    - V = a set of variables
    - E = a set of edge indicating casual relationship
- D-blocked paths
    - Given a path *t*, and a set of nodes *S,*
        - S d-blocks the path *t* if *t* contains:
            - A non-collider which is in S
            - A collider which is not an ancestor of S
- D-seperation
    - http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html#:~:text=d%2Dseparation%20is%20a%20criterion,ness%22%20or%20%22separation%22.
- Markov condition/ assumption
    - Causal graph -> data distribution
- Faithfulness assumption
    - Causal graph <- data distribution
    - Paths cannot cancel out
- Markov equivalence of graph
    - A set of DAGs that are Markov equivalent
    - Completed partial DAG (CPDAG)

| Graphical object | DAG (without hiddens) | MAG | IPG | ADMG (with nested Markov) |
|---|---|---|---|---|
| Type of edges directed / undir. / bidir. / combination | ✓/ - / - / - | ✓/ ✓/ ✓/ - | ✓/ - / ✓/ - | ✓/ - / ✓/ ✓ |
| Correct causal interpretation | ✗ | ✓ | ✓ | ✓ |
| Graphical separation for global Markov | $d$-seperation | $m$-seperation | $m$-seperation | $m$-seperation |
| Criterion for valid adjustment sets | ✓ | ✓ | ? | ✓ |
| Algorithm for identification of intervention distribution | ✓ | ? | ? | ✓ |
| Representation of equivalence class | CPDAG (Markov) | PAG (Markov) | POIPG (Markov) | ? (nested Markov) |
| Independence-based method for learning | PC, IC, SGS | FCI | FCI | - |
| Score-based method for learning | GDS, GES | for linear/binary/ discrete SCMs | ? | for binary/ discrete SCMs |
| Can encode all equality constraints | ✗ | ✗ | ✗ | ✓ (if obs. var. are discrete) |
| Can encode all constraints | ✗ | ✗ | ✗ | ✗ |

Table 9.1: Consider an SCM over (observed) variables $\mathbf{O}$ and (hidden) variables $\mathbf{H}$ that induces a distribution $P_{\mathbf{O},\mathbf{V}}$. How do we model the observed distribution $P_{\mathbf{O}}$? We would like to use an SCM with (arbitrarily many) latent variables. This model class, however, has bad properties for causal learning. This table summarizes some alternative model classes (current research focuses especially on MAGs and ADMGs).

- https://library.oapen.org/bitstream/handle/20.500.12657/26040/11283.pdf?sequ

**Three types of methods**

- **Review of Causal Discovery Methods Based on Graphical Models**
  - https://www.frontiersin.org/articles/10.3389/fgene.2019.00524/full
- Constraint-based
  - PC (Peter Spirtes; Clark Glymour)
    - Assumptions:
      - Markov assumption
      - Faithfulness assumption Acyclicity
      - Causal sufficinecy

Figure 1

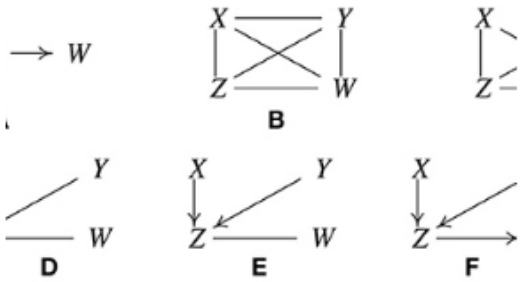


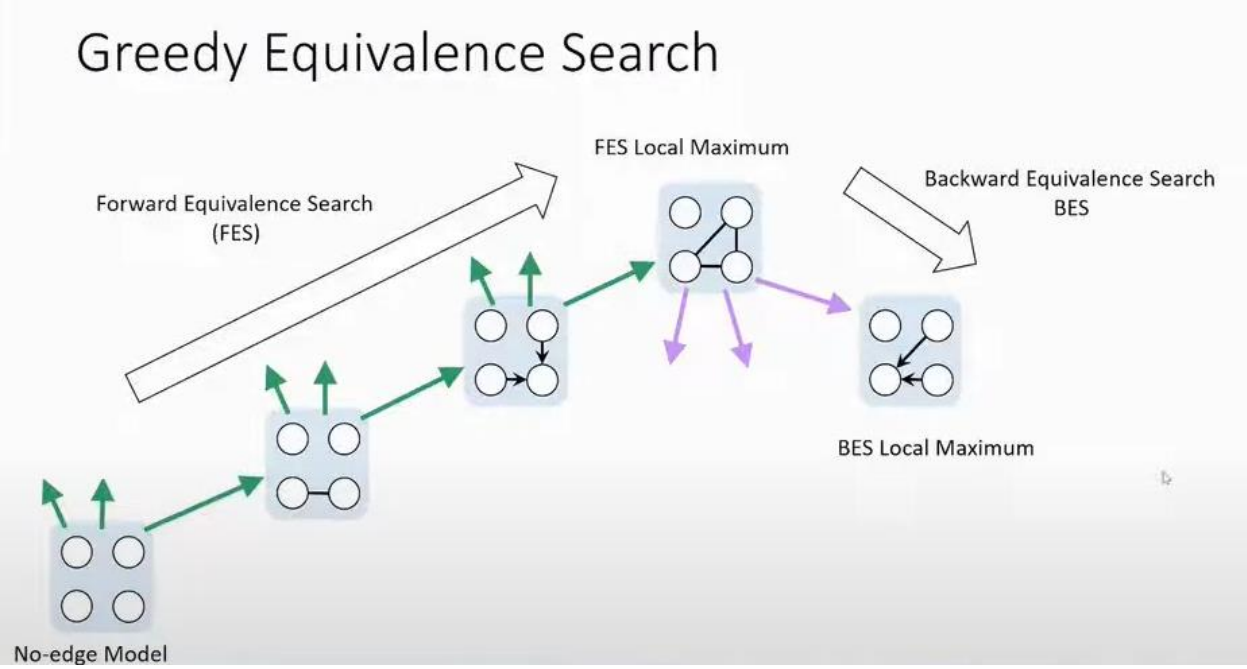

FIGURE 1. Illustration of how the PC algorithm works. (A) Original true causal graph. (B) PC starts with a fully-connected undirected graph. (C) The $X - Y$ edge is removed because $X \perp\!\!\!\perp Y$. (D) The $X - W$ and $Y - W$ edges are removed because $X \perp\!\!\!\perp W \mid Z$ and $Y \perp\!\!\!\perp W \mid Z$. (E) After finding v-structures. (F) After orientation propagation.

- From: https://www.frontiersin.org/articles/10.3389/fgene.2019.00524/full

        - PC returns CPDAG
            - Uncertainty

    - FCI (Fast casual inference)
    - Beyond PC:
        - Relax other assumptions: allow cycles
        - Other difficult situations: e.g. missing data (MVPC, MVFCI)
        - Use any other orientation method (function based etc.)
        - Use any other constrains (pattern-based method)

- Score-based
    - Greedy Equivalent Search (GES)
        - Find graphs best fitting the data
        - BIC
        - Naive score - based methods
            - Super exponential number of graphs with number of data
        - GES
            - **Statistically Efficient Greedy Equivalence Search**
            - http://proceedings.mlr.press/v124/chickering20a/chickering20a.pdf




- Functional Causal Models
    - LinGAM (Linear non-gaussian model)
        - Instead of $$Y = f(X, \varepsilon; \theta_1), \quad (1)$$,

$$Y = bX + \varepsilon, \quad (3)$$

Figure 3

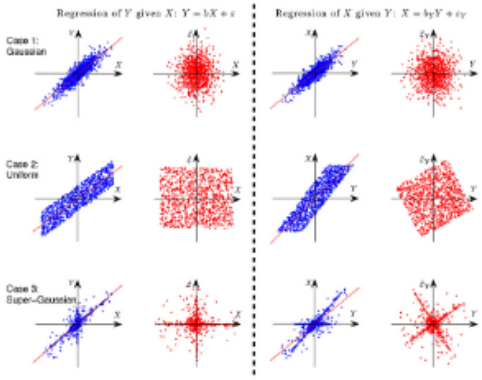


FIGURE 3. Illustration of causal asymmetry between two variables with linear relations. The causal relation is $X \to Y$. From top to bottom: $X$ and $E$ both follow the Gaussian distribution (case 1), uniform distribution (case 2), and Laplace distribution (case 3). The two columns on the left show the scatter plot of $X$ and $Y$ and that of $X$ and the regression residual for regressing $Y$ on $X$, and the two columns on the right correspond to regressing $X$ on $Y$.

        - Find out X causing Y rather than Y causing X
        - Linear Gaussian model is not identifiable
    - Beyond LiNGAM: PNL (Post-linear)
        - **On the Identifiability of the Post-Nonlinear Causal Model**
            - https://arxiv.org/ftp/arxiv/papers/1205/1205.2599.pdf

Table 1: All situations in which the PNL causal model is not identifiable.

| | $p_{e_2}$ | $p_{t_1}$ $(t_1 = g_2^{-1}(x_1))$ | $h = f_1 \circ g_2$ | Remark |
|---|---|---|---|---|
| I | Gaussian | Gaussian | linear | $h_1$ also linear |
| II | log-mix-lin-exp | log-mix-lin-exp | linear | $h_1$ strictly monotonic, and $h_1' \to 0$, as $z_2 \to +\infty$ or as $z_2 \to -\infty$ |
| III | log-mix-lin-exp | one-sided asymptotically exponential (but not log-mix-lin-exp) | $h$ strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$ | — |
| IV | log-mix-lin-exp | generalized mixture of two exponentials | Same as above | — |
| V | generalized mixture of two exponentials | two-sided asymptotically exponential | Same as above | — |

    - Multiple variables: ICA-LINGAM
        - The cause and noise should be independent
            - In matrix form $X = BX + E$
                - $E = (I - B)X$
                - Learn B through ICA
                - $Z = WX$
            - Use Z as E above and we just need to make W have lower triangular shape
            - Linear, Non-Gaussian, Acyclic

**Continuous optimization**

- **Dags with no tears:**
    - https://arxiv.org/pdf/1803.01422.pdf
- Functional Causal Models and Score-based
- Adding DAG constrain

**Causal Discovery + Causal Inference**

- Separate development

        - **Minimal Enumeration of All Possible Total Effects in a Markov Equivalence Class**
            - http://proceedings.mlr.press/v130/guo21c/guo21c.pdf
- Different assumptions

**Deep end-to-end casual inference**

- **Deep End-to-end Causal Inference**
    - https://arxiv.org/pdf/2202.02195.pdf
- Normalizing flow and GNN in deep generative model -> graph and function

**Ongoing research**

- Improved accuracy:
    - Incomplete information
    - accurate casual discovery
- Time-series data adaptation
    - Time dependence noise form
    - Multimodal data

**Scalable Causal AI**

- Casuality:
    - Lots of theory but no impact
- Deep Learning:
    - Great algorithm and impact but not correlation based
- With scalable casual AI