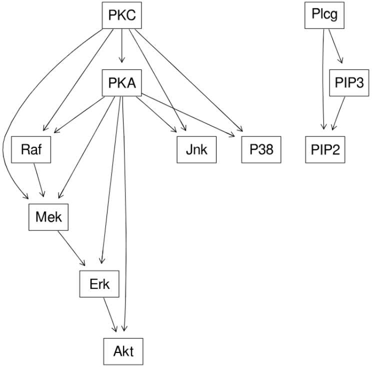
# Expert-In-The-Loop Causal Discovery: Iterative Model Refinement Using Expert Knowledge

Ankur Ankan Johannes Textor

Institute for Computing and Information Sciences
Radboud University, The Netherlands

# What is Causal Discovery?

Data $\Longrightarrow$




Directed Acyclic Graph (DAG)

# Existing Algorithms

| Method | Year | Type |
|---|---|---|
| PC [223] | 1993 | constraint |
| CCD [196] | 1996 | constraint |
| FCI [223] | 2000 | constraint |
| TPDA [31] | 2002 | constraint |
| CPC [190] | 2006 | constraint |
| KCL [227] | 2007 | constraint |
| ION [234] | 2008 | constraint |
| IDA [153] | 2009 | constraint |
| cSAT+ [237] | 2010 | constraint |
| KCI-test [266] | 2012 | constraint |
| RFCI [38] | 2012 | constraint |
| CHC [65] | 2012 | constraint |
| SAT [108] | 2013 | constraint |
| Parallel-PC [141] | 2014 | constraint |
| RPC [89] | 2013 | constraint |
| PC-stable [37] | 2014 | constraint |
| COmbINE [236] | 2015 | constraint |
| backshift [204] | 2015 | - |
| IGSP [256] | 2018 | constraint |
| $\sigma$-CG [57] | 2018 | constraint |
| CCI [225] | 2018 | constraint |
| FCI-soft [132] | 2019 | constraint |
| IBSSI [32] | 2020 | constraint |
| CD-NOD [106] | 2020 | constraint |
| psi-FCI [112] | 2020 | constraint |
| LCDI [264] | 2020 | constraint |
| EG [53] | 2009 | score |
| TWILP [182] | 2014 | score |
| K2 [39] | 1992 | score |
| LB-MDL [140] | 1994 | score |
| HGC [93] | 1995 | score |
| GES [34] | 2002 | score |
| OS [231] | 2005 | score |
| HGL [92] | 2005 | score |
| Meinshausen [159] | 2006 | score |
| Graphical Lasso [60] | 2008 | score |
| BC [9] | 2008 | score |
| TC [185] | 2008 | score |
| HG [91] | 2008 | score |
| Adaptive Lasso [217] | 2010 | score |
| GIES [90] | 2012 | score |
| CD [62] | 2013 | score |
| GBN learner [239] | 2013 | score |
| GES-mod [4] | 2013 | score |
| Pen-PC [87] | 2015 | score |
| Scalable GBN [6] | 2015 | score |
| K-A* [208] | 2016 | score |
| NS-DIST [88] | 2016 | score |
| MIP-GD [181] | 2017 | score |
| CD2 [85] | 2018 | score |
| SP [192] | 2018 | score |
| VAR [261] | 2018 | score |
| GSF [105] | 2018 | score |
| bQCD [229] | 2020 | score |
| GCL [244] | 2020 | score |
| GGIM [56] | 2020 | score |
| GYKZ [69] | 2020 | score |
| SLARAC etc. [252] | 2020 | score |
| Order-MCMC [61] | 2003 | sampling |
| OG [54] | 2008 | sampling |
| EE-DAG [269] | 2011 | sampling |
| ZIPBN [35] | 2020 | sampling |
| LiNGAM [216] | 2006 | asymmetries |
| LV LiNGAM [103] | 2008 | asymmetries |
| non-linear ANM [102] | 2008 | asymmetries |
| CAN [115] | 2009 | asymmetries |
| CCM [226] | 2012 | asymmetries |
| IGCI [114] | 2012 | asymmetries |
| KCDC [161] | 2018 | asymmetries |
| MMHC [238] | 2006 | hybrid |
| ARGES [171] | 2018 | hybrid |

| Method | Year | Data |
|---|---|---|
| CMS [152] | 2014 | low |
| NO TEARS [267] | 2018 | low |
| CGNN [75] | 2018 | low |
| Graphite [83] | 2019 | low/medium |
| SAM [122] | 2019 | low/medium |
| DAG-GNN [262] | 2019 | low |
| GAE [177] | 2019 | low |
| NO BEARS [142] | 2019 | low/medium/high |
| Meta-Transfer [19] | 2019 | Bi |
| DEAR [214] | 2020 | high |
| CAN [167] | 2020 | low/medium/high |
| NO FEARS [251] | 2020 | low |
| GOLEM [176] | 2020 | low |
| ABIC [20] | 2020 | low |
| DYNOTEARS [178] | 2020 | low |
| SDI [124] | 2020 | low |
| AEQ [64] | 2020 | Bi |
| RL-BIC [272] | 2020 | low |
| CRN [125] | 2020 | low |
| ACD [151] | 2020 | low |
| V-CDN [145] | 2020 | high |
| CASTLE (reg.) [138] | 2020 | low/medium |
| GranDAG [139] | 2020 | low |
| MaskedNN [175] | 2020 | low |
| CausalVAE [257] | 2020 | high |
| CAREFL [126] | 2020 | low |
| Varando [244] | 2020 | low |
| NO TEARS+ [268] | 2020 | low |
| ICL [250] | 2020 | low |
| LEAST [271] | 2020 | low/medium/high |

List of causal discovery algorithms [1]

[1] Vowels, Matthew J., Necati Cihan Camgoz, and Richard Bowden. "D'ya like dags? A survey on structure learning and causal discovery."

# Causal Discovery in Practice

- However, in applied research their adoption is limited.
- Tennant et al. (2021)[2]: Reviewed 234 papers in health science reporting DAGs.
  - None employed causal discovery methods.
- Petersen et al. (2021)[3]:
  "Although causal discovery algorithms have been available for a long time, their use in epidemiology is limited to only a few studies"

[2] Tennant, et al. "Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations." International Journal of Epidemiology.

[3] Petersen, et al. "Data-driven model building for life-course epidemiology." American Journal of Epidemiology"

# Causal Discovery in Practice

Gap in causal discovery method development and their application.

- **Lack of Trust:**
  - Most algorithms are asymptotically consistent.
  - But finite sample properties are not well understood.
  - May produce outputs that contradict domain knowledge.
  - Lack of performance evaluation methods.
- **Outputs Markov Equivalence Class:**
  - Multiple DAGs are consistent with an observational dataset.
  - Algorithms can only recover the Markov Equivalence Classes.
  - Downstream tasks like identification, and effect estimation typically require a fully specified DAG.

As a result, practitioners mostly draw DAGs using only domain knowledge.

# Expert-In-The-Loop Causal Discovery

- Expert-In-The-Loop Causal Discovery tries to address this gap.
- We believe domain experts:
  - are often good at determining causal directions, i.e., ancestral relationships.
  - may sometimes struggle to identify absence of casual links.
  - can find it difficult to distinguish direct from indirect effects.
- Expert-In-The-Loop assists practitioners in constructing DAGs:
  - Suggests addition or removal of edges.
  - Lets the expert choose and specify ancestral relationships.
  - Keeps the expert in control of the model building process.

# Algorithm: Expert-In-The-Loop

```
Function ExpertInLoop(V, 𝒟, 𝒜):
    E_p ← ∅ /* Current edges */
    B ← ∅ /* Edges that were pruned or removed from cycle */
    repeat
        E ← E_p
        (V, E, B) ← Expand(V, E, 𝒟, 𝒜, B, 1)
        (V, E_p) ← Prune(V, E, 𝒟)
        B ← B ∪ {E \ E_p}
    until E = E_p
    return (V,E)
```

# Algorithm: Expand

**Function** $\textsc{Expand}(V, E, \mathcal{D}, \mathcal{A}, B, k)$:

$L \leftarrow \{\}$

**foreach** $X, Y$ *where* $X \rightarrow Y \notin E \cup B$ *and* $Y \rightarrow X \notin E \cup B$ **do**

  $\mathbf{Z}$ be a set that $d$-separates $X$ and $Y$ in $(V, E)$

  **if** $\mathcal{D}(X, Y, \mathbf{Z}) = 0$ **then**

    $L \leftarrow L \cup \mathcal{A}(X, Y)$

  **end**

  **if** $|L| \geq k$ **then**

    **go to** 12

  **end**

**end**

$R \leftarrow \textsc{FixCycles}(V, E \cup L, \mathcal{D})$

$B \leftarrow B \cup R$ ; $E \leftarrow (E \cup L) \setminus R$

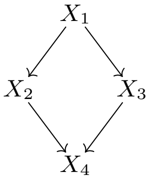
**return** $(V, E, B)$

# Algorithm: Prune

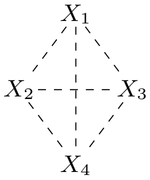
**Function** $\textsc{Prune}(V, E, \mathcal{D})$**:**

```
R ← {}
foreach X → Y ∈ E do
    let Z be a set that d-separates X and Y in (V, E \ {X → Y})
    if D(X, Y, Z) = 1 then
        R ← R ∪ {X → Y}
    end
end
E ← E \ R
return (V, E)
```

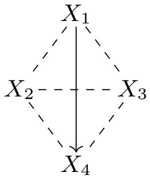
# Examples



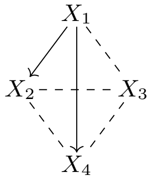

(a) True DAG



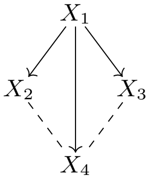

(b) No edges.



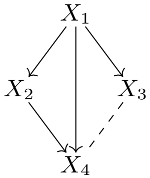

(c) $X_1 \to X_4$



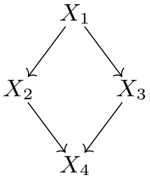

(d) $X_1 \to X_2$




(e) $X_1 \to X_3$




(f) $X_2 \to X_4$




(g) $X_3 \to X_4$, $X_1 \not\to X_4$

# Theoretical Analysis: Expert-In-The-Loop

- For theoretical analysis, we assume d-separation oracle and an expert oracle.
- We consider two types of expert oracles.
  - **Strong Oracle:** Always gives the correct ancestral relationship including if there is no ancestral relationship.
  - **Weak Oracle:** Gives correct answers when an ancestral relationship exists; otherwise can give an incorrect answer.
- Both of these oracles are able to recover the correct DAG.

# Empirical Analysis: Comparison with Other Algorithms

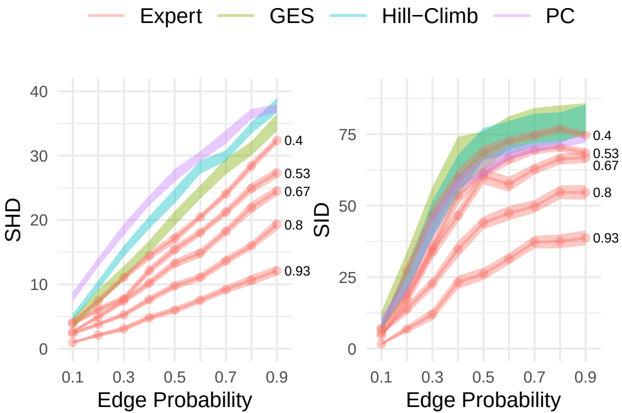
- Simulated linear Gaussian data from random DAGs.
- Simulated expert responses with accuracy $\alpha$ as:

$$x = \mathrm{rand}([0,1])$$

$$\mathrm{Expert}(\alpha) = \begin{cases} \text{True Ancestral Reln.}, & \text{if } x <= \alpha \\ \mathrm{rand}(X \rightarrow Y, Y \leftarrow X, \mathrm{None}) & \text{otherwise} \end{cases}$$

- Compared the SHD and SID with other algorithms.

# Empirical Analysis: Comparison with Other Algorithms




Comparison of PC, Hill-Climb Search, and GES algorithms with ExpertInLoop algorithm with varying values of expert accuracy, $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

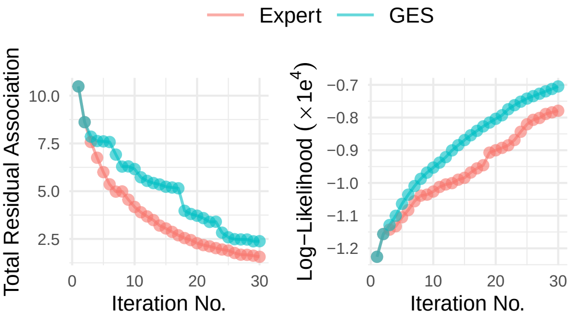
# Total Residual Association

- Total Residual Association: An absolute measure of DAG fit, $\tau$:

$$\tau = \sum_{\substack{X,Y \in V \\ X \to Y, Y \to X \notin E}} \phi(X, Y, \mathrm{pa}_G(X) \cup \mathrm{pa}_G(Y))$$

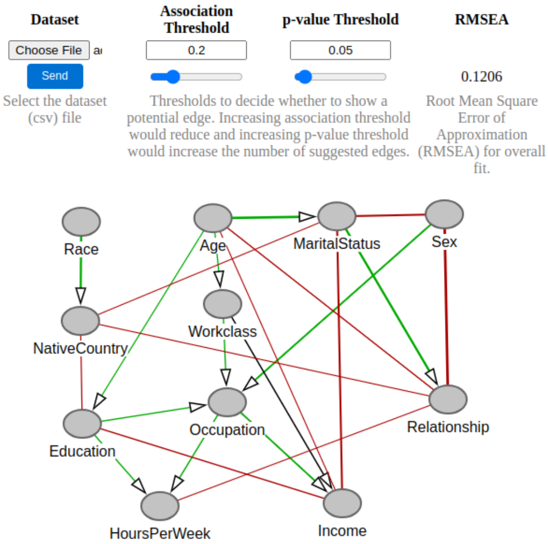
  where $\phi(X, Y, \mathbf{Z})$ is the effect size of a conditional independence test $X \perp Y | \mathbf{Z}$
- $\tau$ approaches 0 with improving model fit.
- Unlike likelihood based measures, $\tau$ can be used to validate the fit of the DAG.

# Empirical Analysis: Total Residual Association




Expert-In-the-Loop causal discovery on the Adult Income dataset. Two measures of fit are shown over 30 iterative modifications: total residual association (lower is better); log-likelihood (right, higher is better).

# Practical Implementations: Web-Based Tool



# Practical Implementations: Python Implementation

```python
from pgmpy.estimators import ExpertInLoop
from pgmpy.utils import llm_pairwise_orient

descriptions = {
    "Age": "The age of a person",
    "Workclass": "The workplace where the person is ...",
    "Education": "The highest level of education the ...",
    "MaritalStatus": "The marital status of the person",
    "Occupation": "The kind of job the person does.",
    "Relationship": "The relationship status of the person",
    "Race": "The ethnicity of the person",
    ...
}

dag = ExpertInLoop(data).estimate(
    variable_descriptions=descriptions,
    orientation_fn=llm_pairwise_orient,
    llm_model="gemini/gemini-1.5-flash",
    pval_threshold=0.05,
    effect_size_threshold=0.05
)
```

# Conclusions and Future Work

- Expert-In-The-Loop: An iterative and interactive approach to assist in constructing DAGs.
- Outperforms fully automated algorithms if the expert correctly orients edges in at least two-thirds of cases.
- Propose Total Residual Association as a measure to validate the fit of a DAG.
- Iterative improvement risks overfitting.
- The assumption of no unobserved confounding is restrictive.