# EMPIRICAL BAYESIAN APPROACHES FOR ROBUST CONSTRAINT-BASED CAUSAL DISCOVERY UNDER INSUFFICIENT DATA

Zijun Cui (presenter) [1]
cuiz3@rpi.edu

Naiyu Yin [1]

Yuru Wang [2]

Qiang Ji [1]
jiq@rpi.edu

1 RENSSELAER POLYTECHNIC INSTITUTE

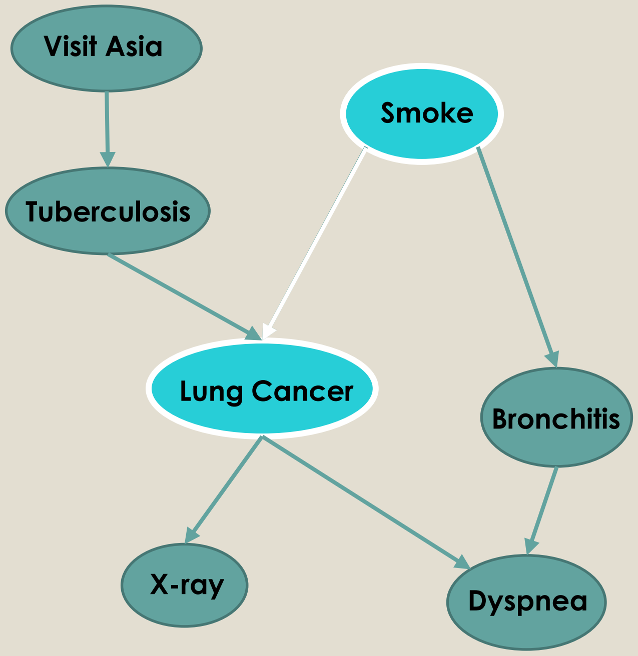
2 NORTHEAST NORMAL UNIVERSITY

# Causal Discovery

- Causal relations among variables are captured by a directed acyclic graph (DAG)
  - A direct link from node $X$ to node $Y$ indicates the cause-effect relation between cause variable $X$ and effect variable $Y$

- Causal discovery is to learn a DAG capturing cause-effect relationships among a set of random variables from observational data

- **Causal discovery under insufficient data is of great importance**
  - Existing methods are focused on learning a DAG with high confidence under sufficient data
  - However, in many domains, the availability of data is very limited



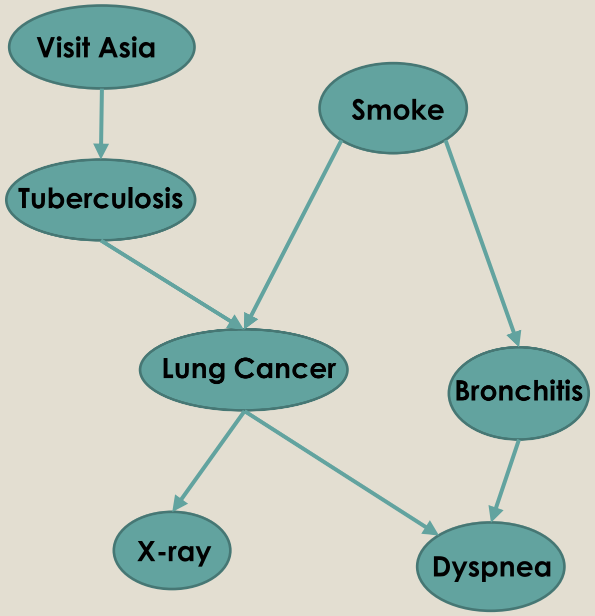

#2328

Example structure is revised based on ASIA dataset in Bnlearn Repository

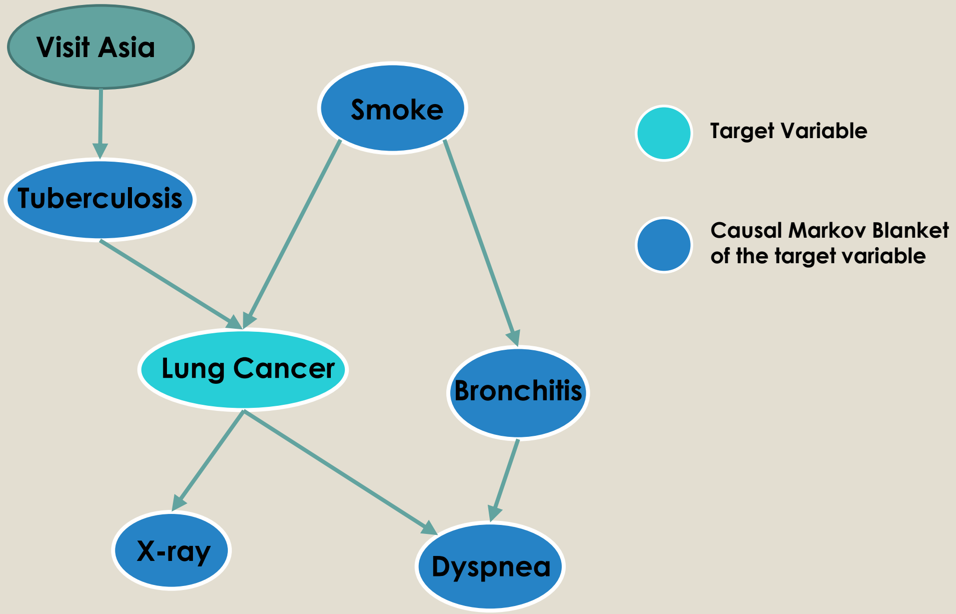
# Constraint-based Causal Discovery

#2328

- Constraint-based causal discovery methods apply independence tests to determine a DAG from observational data
- It can be performed globally or locally




**Global approaches** aim at learning cause-effect relationships among all random variables




**Local approaches** aim at identifying the direct causes and effects of a target variable, represented by a causal Markov blanket

Example structure is revised based on ASIA dataset in Bnlearn Repository

# Bayesian Approaches for Independence Tests



- For both global and local approaches, the main challenge of the constraint-based causal discovery is that its performance highly depends on the accuracy of the independence test

- We propose two Bayesian-augmented frequentist independence tests
  - Bayesian approach is adopted to reliably estimate independence test statistics with limited data by considering the entire parameter space instead of using a point estimate one
  - The Bayesian statistics are then used by frequentist independence tests

- Specifically, we introduce Bayesian approach for two types of independence tests
  - mutual Information based independence test
  - statistical testing based independence test

# Independence Test

IJCAI ECAI VIENNA 22

#2328

- **Mutual information based independence test**
  - The mutual information (MI) of two discrete random variables $X$ and $Y$ is defined as

$$\mathrm{MI}(X;Y) = \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} \theta_{ij} \ln \frac{\theta_{ij}}{\theta_i \theta_j}$$

$K_x$ and $K_y$ denote the total number of possible states of $X$ and $Y$. $\theta_i = p(x_i), \theta_j = p(y_j)$ and $\theta_{ij} = p(x_i, y_j)$ are probability distribution parameters

  - If $\mathrm{MI}(X;Y) < \mathrm{Threshold}$, $X$ and $Y$ are declared to be independent; Otherwise, $X$ and $Y$ are dependent.

- **Statistical testing based independence test**
  - G-test is a standard likelihood ratio test. Its statistics $g$ asymptomatically follows the $\chi^2_{df=(K_x-1)(K_y-1)}$ distribution and is defined as

$$g = -2 \sum_{i=1}^{K_x} \sum_{j=1}^{K_y} n_{ij} \ln \frac{\theta_i \theta_j}{\theta_{ij}}$$

  - If $p$-value is smaller than the significance level (default 5%), the null hypothesis is rejected and the alternative hypothesis is accepted. Thus, $X$ and $Y$ are declared to be dependent; Otherwise, $X$ and $Y$ are declared to be independent.

- **Independence Test Accuracy under insufficient data**
  - Existing methods perform a Maximum Likelihood estimation (MLE) of the parameters $\boldsymbol{\theta}$ directly from data $D$, i.e.,

$$\boldsymbol{\theta} = \mathrm{argmax}\, P(D|\boldsymbol{\theta})$$

  - The MLE estimates are inaccurate when $D$ is insufficient. As a result, independence tests are subject to errors under limited data

# Bayesian Approach for Mutual Information based Independence Test

#2328

- Full Bayesian MI is based on estimating expected MI over data $D$ :

$$MI^{FB}(X;Y|D) = \int\int MI\,(X;Y|\boldsymbol{\theta})p(\boldsymbol{\theta},\alpha|D)d\boldsymbol{\theta}d\alpha = \int\int MI\,(X;Y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha,D)p(\alpha|D)d\boldsymbol{\theta}d\alpha$$

- The integration over $\alpha$ is approximated by maximizing it out as

$$MI^{eB}(X;Y|D) = \int\int MI\,(X;Y|\boldsymbol{\theta})p(\boldsymbol{\theta},\alpha|D)d\boldsymbol{\theta}d\alpha = \int MI\,(X;Y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha^*,D)d\boldsymbol{\theta}$$

with $\alpha^* = \text{argmax}\ p(\alpha|D) = \text{argmax}\ p(D|\alpha)p(\alpha)$. Assuming $p(\alpha)$ follows the uniform distribution, we have $\alpha^* = \text{argmax}$ $p(D|\alpha)$ and can be solved through a fixed-point update

- Given the $\alpha^*$, we in the end have

$$MI^{eB}(X;Y|D) = \psi(N+\alpha^*K+1) - \sum_{ij}\frac{n_{ij}+\alpha^*}{N+\alpha^*K}[\psi\left(n_i+\alpha^*K_y+1\right)+\psi\left(n_j+\alpha^*K_x+1\right)-\psi\left(n_{ij}+\alpha^*+1\right)]$$

where $\psi(x)$ is the digamma function. $n_i$ and $n_j$ are the number of samples for $X=i$ and $Y=j$ respectively, and $n_{ij}$ is the number of samples for $(X,Y)=(i,j)$

# Bayesian Approach for Statistical Testing based Independence Test

#2328

- A Bayesian estimate of hypothesis likelihood is considered as

$$BF = \frac{P(D|H_0,\alpha_0)}{P(D|H_1,\alpha_1)} = \frac{\int P(D|\boldsymbol{\theta}, H_0)P(\boldsymbol{\theta}|H_0, \alpha_0)d\boldsymbol{\theta}}{\int P(D|\boldsymbol{\theta}, H_1)P(\boldsymbol{\theta}|H_1, \alpha_1)d\boldsymbol{\theta}}$$

$\alpha_0$ and $\alpha_1$ are the respective hyper-parameters under null and alternative hypothesis

- To apply BF for a statistical testing, like $G$ test, we approximate it as

$$\widetilde{BF} = \frac{P(D|H_0,\widetilde{\boldsymbol{\theta}})}{P(D|H_1,\widetilde{\boldsymbol{\theta}})} = \frac{\prod_{i=1}^{K_x} \tilde{\theta}_i^{n_i} \prod_{j=1}^{K_y} \tilde{\theta}_j^{n_j}}{\prod_{i=1,j=1}^{K_x,K_y} \tilde{\theta}_{ij}^{n_{ij}}}$$

with $\tilde{\theta}_k = \frac{a^* n_k + b^* \alpha}{a^* N + b^* K\alpha}$ and $\Lambda = \binom{a}{b}$ are unknown coefficients that can be solved analytically

- The statistic $BF_{chi2}$ in the end is computed as

$$BF_{chi2} = -2\ln\widetilde{BF} = -2\sum_{i=1}^{K_x}\sum_{j=1}^{K_y} n_{ij}\ln\frac{\tilde{\theta}_i\tilde{\theta}_j}{\tilde{\theta}_{ij}}$$

$BF_{chi2}$ asymptomatically follows the distribution $\chi^2_{df=(K_x-1)(K_y-1)}$. We set 5% as the default significance level

# Local Causal Discovery

IJCAI ECAI VIENNA 22

#2328

- We consider the causal Markov blanket (CMB) for comparison
- $cI^{eB}$ denotes the CMB with empirical Bayesian MI estimation; $cBF_{chi2}$ denotes the CMB with $BF_{chi2}$ independence test

| Dataset | Size | SHD | | | #Independence Test | | |
|---|---|---|---|---|---|---|---|
| | | $cI^{eB}$ | $cBF_{chi2}$ | CMB | $cI^{eB}$ | $cBF_{chi2}$ | CMB |
| CHILD | 100 | 2.90±0.28 | 2.65±0.40 | 5.94±0.65 | 1008 | 1154 | 16869 |
| | 300 | 2.61±0.26 | 2.64±0.59 | 6.95±0.63 | 1709 | 1926 | 14578 |
| | 500 | 2.29±0.31 | 2.24±0.84 | 4.52±0.58 | 2524 | 4751 | 13873 |
| | **MEAN** | 2.60 | **2.51** | 5.80 | **1747** | 2610 | 15107 |
| INSURANCE | 100 | 3.89±0.34 | 3.98±0.39 | 7.18±0.66 | 1261 | 1363 | 22168 |
| | 300 | 3.47±0.21 | 3.24±0.12 | 7.59±0.57 | 1541 | 2977 | 18043 |
| | 500 | 3.11±0.21 | 2.98±0.13 | 7.20±0.67 | 1477 | 3949 | 14881 |
| | **MEAN** | 3.49 | **3.40** | 7.32 | **1426** | 2763 | 18364 |
| ALARM | 100 | 2.69±0.07 | 2.39±0.19 | 5.20±0.71 | 1424 | 1109 | 27492 |
| | 300 | 2.50±0.19 | 2.27±0.15 | 4.36±0.83 | 2398 | 3885 | 14900 |
| | 500 | 2.40±0.11 | 2.26±0.19 | 3.53±0.62 | 2807 | 4766 | 11328 |
| | **MEAN** | 2.53 | **2.31** | 4.36 | **2210** | 3253 | 17907 |
| HAILFINDER | 500 | 3.33±0.02 | 4.22±0.04 | 7.90±0.11 | 676 | 1923 | 183350 |
| | 800 | 3.56±0.01 | 4.49±0.13 | 7.12±0.09 | 1098 | 2145 | 169705 |
| | 1000 | 3.56±0.09 | 4.45±0.08 | 7.10±0.11 | 1924 | 2621 | 119815 |
| | **MEAN** | **3.48** | 4.39 | 7.37 | **1233** | 2229 | 157620 |
| CHILD3 | 500 | 2.46±0.23 | 2.53±0.18 | 4.72±0.28 | 7168 | 7417 | 14789 |
| | 800 | 3.01±0.13 | 2.67±0.11 | 3.57±0.21 | 6720 | 7802 | 9765 |
| | 1000 | 2.90±0.07 | 2.57±0.23 | 3.09±0.19 | 8424 | 8285 | 9516 |
| | **MEAN** | 2.79 | **2.59** | 3.79 | **7437** | 7835 | 11357 |
| CHILD5 | 500 | 2.87±0.05 | 2.62±0.19 | 5.00±0.15 | 5234 | 11126 | 16819 |
| | 800 | 2.66±0.21 | 3.02±0.13 | 5.75±0.32 | 8236 | 11424 | 51967 |
| | 1000 | 2.82±0.23 | 2.99±0.07 | 4.34±0.19 | 13384 | 9956 | 36888 |
| | **MEAN** | **2.78** | 2.88 | 5.03 | **8951** | 10835 | 26322 |

- Both $cI^{eB}$ and $cBF_{chi2}$ outperform CMB in terms of both accuracy (SHD) and efficiency (# Independence Test)
- Comparing the performance between the two proposed methods
  - $cBF_{chi2}$ achieves overall better accuracy
  - $cI^{eB}$ is more efficient with the fewest number of independence tests on all datasets

# Global Causal Discovery

#2328

- We consider the RAI-BF and PC-Stable for comparison
- $rI^{eB}$ denotes the RAI with empirical Bayesian MI estimation; $rBF_{chi2}$ denotes the RAI with $BF_{chi2}$ independence test

| | | SHD | | | | #Independence Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | Size | $rI^{eB}$ | $rBF_{chi2}$ | RAI-BF | PC-Stable | $rI^{eB}$ | $rBF_{chi2}$ | RAI-BF | PC-Stable |
| CHILD | 100 | 21.6±2.1 | 24.2±2.3 | 30.4±3.7 | 23.8±1.7 | 283 | 314 | 893 | 559 |
| | 300 | 19.9±2.7 | 17.7±1.8 | 23.5±4.4 | 22.6±1.9 | 342 | 546 | 997 | 986 |
| | 500 | 17.6±1.7 | 16.0±2.9 | 22.6±2.4 | 24.4±2.2 | 424 | 754 | 975 | 1317 |
| | **MEAN** | 19.7 | **19.3** | 25.5 | 23.6 | **350** | 538 | 955 | 954 |
| INSURANCE | 100 | 48.9±1.3 | 50.1±2.9 | 54.9±3.6 | 52.0 ±1.5 | 486 | 604 | 905 | 1217 |
| | 300 | 47.3±0.8 | 44.5±2.0 | 46.6±3.2 | 50.2±3.1 | 576 | 986 | 1011 | 1250 |
| | 500 | 49.5±1.8 | 39.4±3.0 | 47.1±2.2 | 50.7±2.5 | 662 | 1200 | 1120 | 2326 |
| | **MEAN** | 48.6 | **44.7** | 49.5 | 51.0 | **575** | 930 | 1012 | 1598 |
| ALARM | 100 | 44.5±2.2 | 42.7±2.3 | 48.4±5.8 | 45.8 ±4.9 | 891 | 958 | 1591 | 2215 |
| | 300 | 40.7±3.0 | 36.1±4.5 | 35.3±5.4 | 34.6±2.7 | 1158 | 1752 | 1881 | 3398 |
| | 500 | 40.0±3.1 | 29.8±5.1 | 29.8±5.2 | 36.5 ±5.7 | 1433 | 2018 | 2098 | 3992 |
| | **MEAN** | 41.7 | **36.2** | 37.8 | 39.0 | **1161** | 1576 | 1857 | 3202 |
| HAILFINDER | 500 | 88.0±2.0 | 98.3 ±1.5 | 118.0±1.0 | 91.6±1.0 | 2024 | 2587 | 6171 | 3267 |
| | 800 | 85.0±1.7 | 106.3 ±2.1 | 124.7 ±6.7 | 99.7±1.2 | 1983 | 3726 | 7847 | 3423 |
| | 1000 | 92.3±4.5 | 108.3 ±2.3 | 131.3 ±3.2 | 101.8±2.2 | 2638 | 3073 | 16618 | 3603 |
| | **MEAN** | **88.4** | 104.3 | 124.7 | 97.7 | **2215** | 3129 | 10212 | 3431 |
| CHILD3 | 500 | 67.6±3.2 | 54.3±2.6 | 79.6±4.9 | 81.2±2.8 | 2693 | 3796 | 5422 | 4963 |
| | 800 | 65.8±2.5 | 52.9±2.8 | 74.0±3.7 | 79.9±2.4 | 3941 | 4587 | 5106 | 6026 |
| | 1000 | 61.5±3.8 | 52.3±3.9 | 71.0±6.5 | 81.4±2.7 | 4723 | 5170 | 5980 | 6846 |
| | **MEAN** | 65.0 | **53.2** | 74.9 | 80.8 | **3786** | 4518 | 5503 | 5945 |
| CHILD5 | 500 | 122.0±2.6 | 109.3±5.1 | 134.0 ±2.6 | 113.9±2.4 | 6966 | 8646 | 10038 | 10253 |
| | 800 | 121.7±3.8 | 105.3±4.0 | 132.3±6.7 | 120.1±2.9 | 10249 | 10431 | 9337 | 10708 |
| | 1000 | 116.3±2.9 | 105.7±2.5 | 126.3±7.0 | 123.4±1.7 | 10375 | 10494 | 11174 | 11070 |
| | **MEAN** | 120.0 | **106.8** | 126.3 | 119.1 | **9197** | 9857 | 11174 | 10677 |

- Both $rI^{eB}$ and $rBF_{chi2}$ outperform RAI-BF and PC-Stable in terms of both accuracy (SHD) and efficiency (# Independence Test)
- Comparing the performance between the two proposed methods
  - $rBF_{chi2}$ achieves overall better accuracy
  - $rI^{eB}$ achieves overall better efficiency
- We reach consistent conclusions

# Conclusions

- We introduce Bayesian methods for robust constraint-based causal discovery under insufficient data
- Two Bayesian-augmented frequentist independence tests are proposed for reliable statistic estimation under a frequentist independence test framework
- Through extensive experiments, we show that, by introducing Bayesian approaches, the proposed methods not only outperform the competing methods in terms of accuracy, but also improve efficiency significantly

# Thank You!