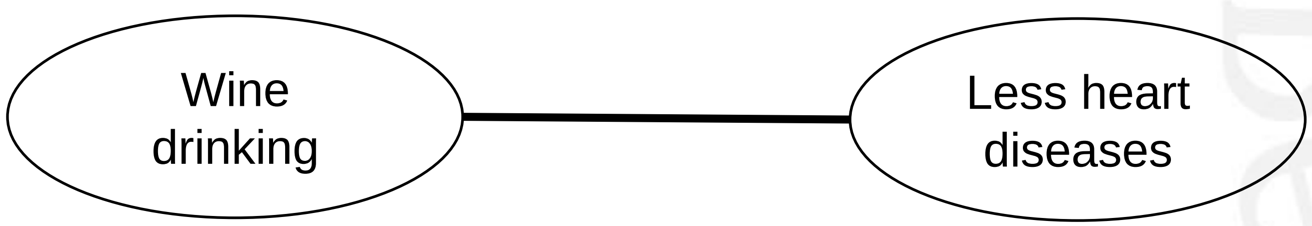
Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD

Elena Sokolova, Daniel von Rhein, Jilly Naaijen,
Perry Groot, Tom Claassen, Jan Buitelaar and Tom Heskes
Radboud University, Nijmegen The Netherlands

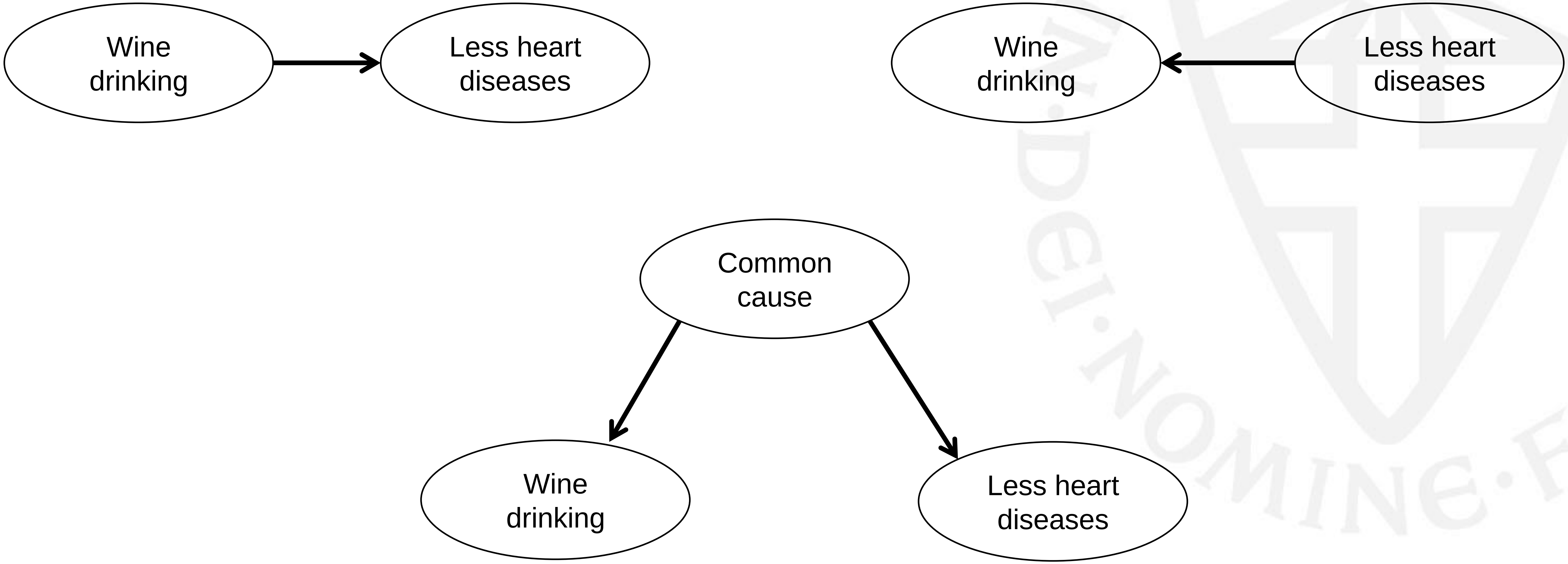
Does wine drinking prevent heart disease?

Wine drinking and lower rate of heart disease are associated



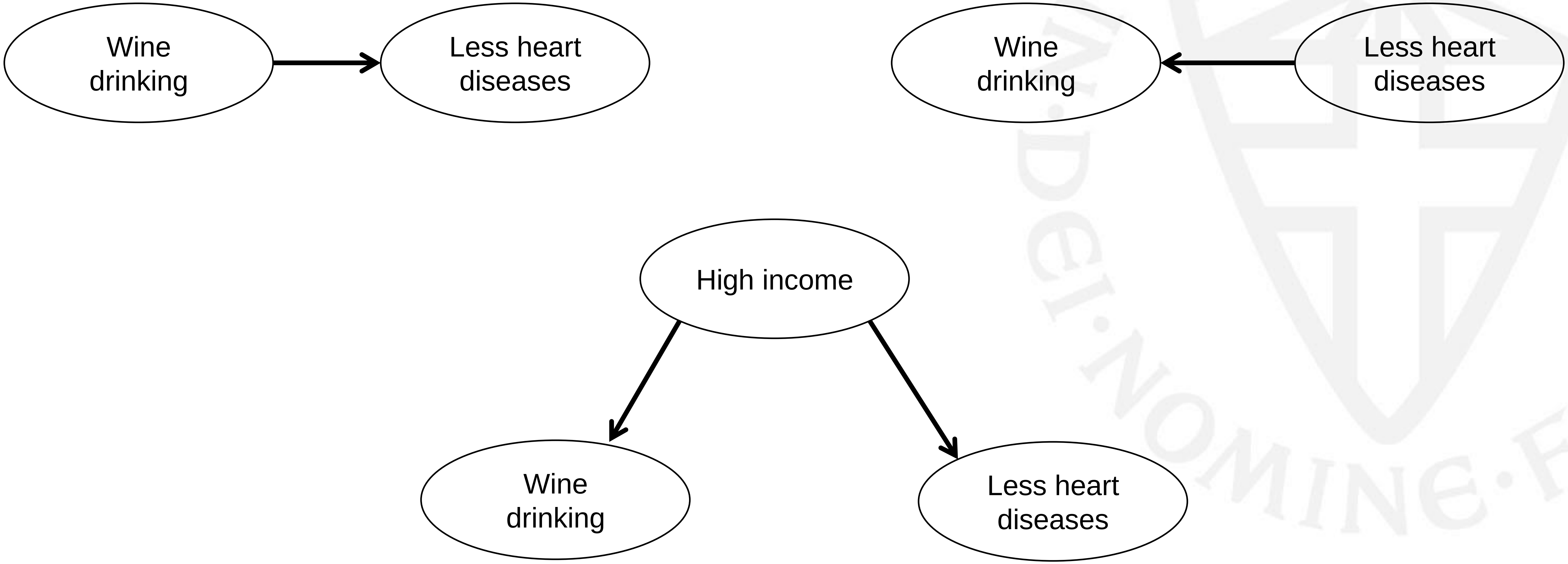
Does wine drinking prevent heart disease?

All possible models



Does wine drinking prevent heart disease?

All possible models



A way to learn causality

1. Take randomly 200 people
2. Randomly split them in **controls** and **treatment** groups
3. Force **treatment** group to drink wine, forbid **control** group to drink wine
4. Wait 40 years
5. Measure correlation

[Randomized Controlled Trial]

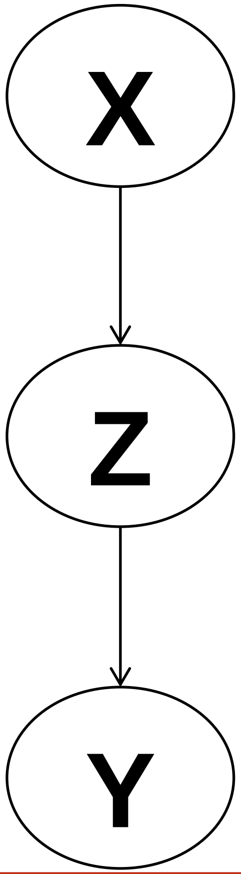
Can we learn causal relationships from observed data?

Yes!

Conditional Independence

X and ***Y*** are conditionally independent given ***Z*** :
Given ***Z***

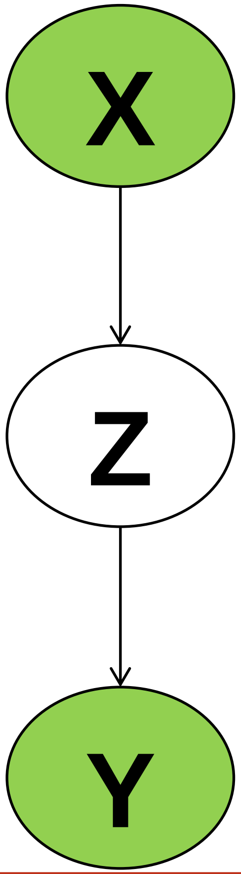
- knowledge of ***X*** provides no information for ***Y***
- knowledge of ***Y*** provides no information for ***X***



Conditional Independence

X and ***Y*** are conditionally independent given ***Z*** :
Given ***Z***

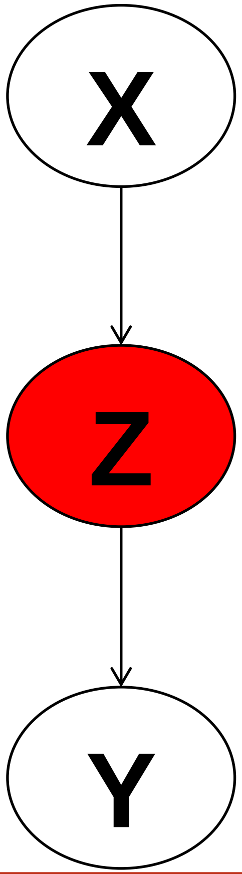
- knowledge of ***X*** provides no information for ***Y***
- knowledge of ***Y*** provides no information for ***X***



Conditional Independence

X and ***Y*** are conditionally independent given ***Z*** :
Given ***Z***

- knowledge of ***X*** provides no information for ***Y***
- knowledge of ***Y*** provides no information for ***X***



Learning causal network

Bayesian constraint-based causal discovery:

- Uses Bayesian approach to estimate the reliability of the causal statements, avoiding propagation of unreliable decisions

T. Claassen, T. Heskes. **A Bayesian approach to constraint based causal inference**. In *UAI 2012*

BCCD

Basic idea:

- **Step 0** Start with a fully connected graph.
- **Step 1** Estimate the reliability of a causal statement $(X \rightarrow Y)$ using Bayesian score.
- **Step 2** If a causal statement declares a variable conditionally independent, delete an edge.
- **Step 3** Rank all causal statements and orient edges in the graph.

BCCD

The reliability of the causal statement L given the data **D** using Bayesian score**:**

$$p(L|D) = \frac{\sum_{\mathcal{M} \in M(L)} p(D|\mathcal{M})p(\mathcal{M})}{\sum_{\mathcal{M} \in M} p(D|\mathcal{M})p(\mathcal{M})}$$

There is a closed form solution for $p(D|\mathcal{M})$:

- Discrete random variables - BD metric
- Continuous Gaussian variables - BGe metric

BCCD

Advantages of the method:

- Robust
- Can handle latent variables
- Gives an indication whether an edge does exist or not

Limitation of the method:

- Works only with discrete variables or Gaussian variables
- Cannot handle missing values

Undirected graphs

- Precision matrix- inverse of correlation matrix
- Precision matrix - the set of conditional independencies
- Add sparsity constraints

Undirected graphs

Glasso to find optimum

$$\Theta_\lambda = \text{argmax}_\Theta \{\underbrace{\text{logdet}(\Theta) - \text{tr}(\Theta S)}_{\text{Goodness of fit}} - \underbrace{\lambda\|\Theta\|_1}_{\text{Sparsity penalty}}\}$$

- $\Theta = \Sigma^{-1}$ inverse of correlation matrix
- S- empirical correlation matrix

- Spearman instead of Pearson partial correlation
- Adjust Spearman correlation, to make it closer to Pearson
- Shift correlation matrix to the closest one if it is negative definite
- Use EM if there are missing values

Assumptions

- Data is a mixture of discrete and continuous variables
- Data is missing completely at random (MCR)
- Relationships between variables are monotonic, i.e. variables follow a so-called non paranormal distribution

Method extension

- BIC score:

$$BIC\ score(\boldsymbol{D}|\mathcal{G}) = \underbrace{M\sum_{i=1}^{n} I(X_i, Pa_{X_i})}_{\text{Goodness of fit}} - \underbrace{\frac{\log M}{2}\mathrm{Dim}[\mathcal{G}]}_{\text{Complexity penalty}}$$

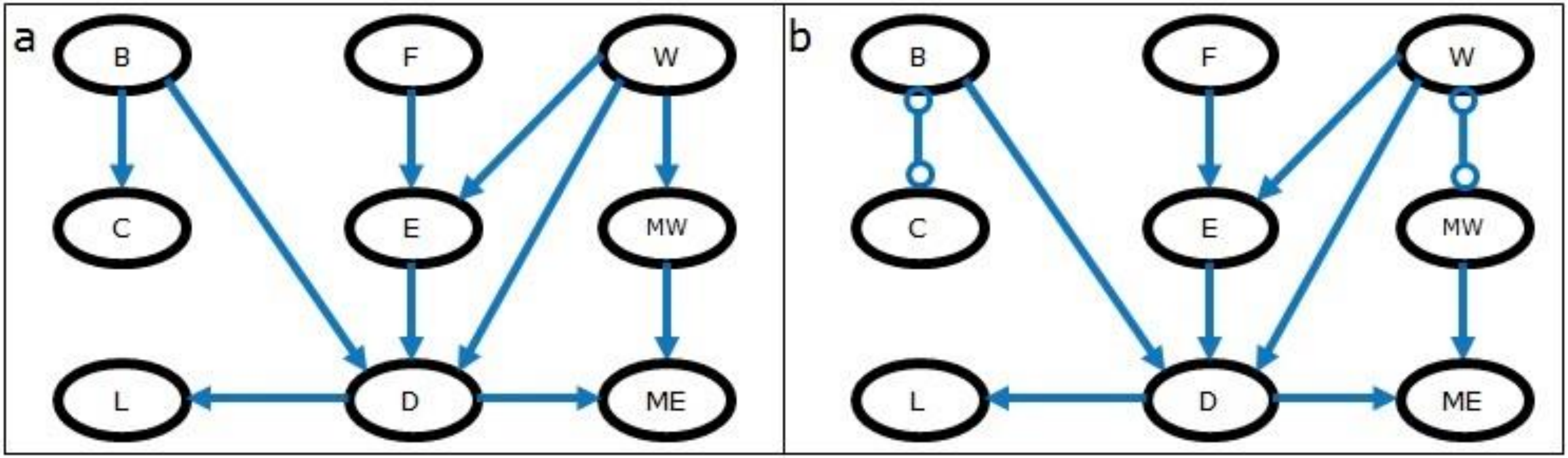
- Mutual information

$$I(x_1, \dots, x_n) = -\frac{1}{2}\log\frac{|R|}{|R_{Pa_i}|}$$

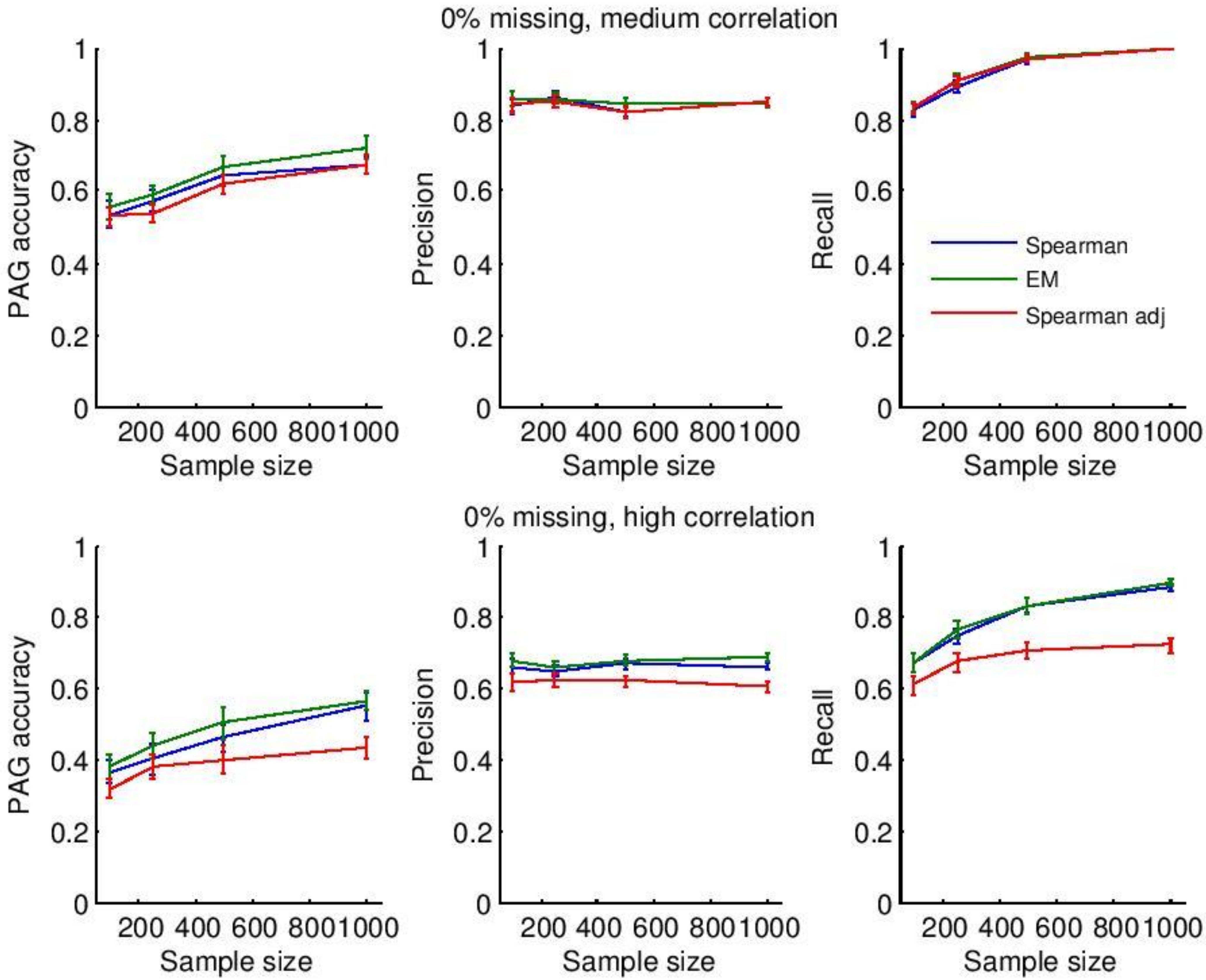
- Use Spearman instead of Pearson
- Use EM if there are missing values

Simulated data

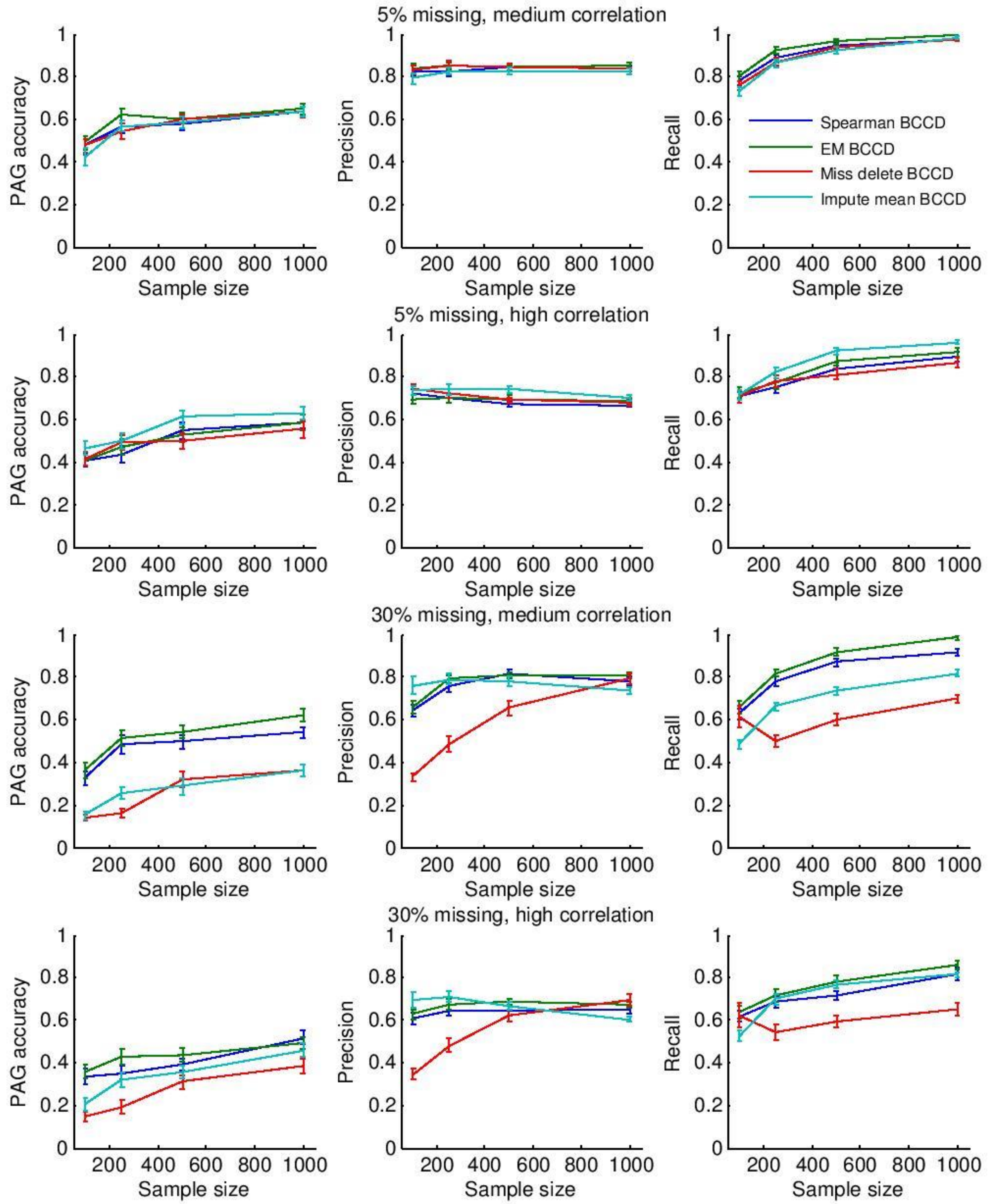
- Waste Incinerator Network, x^3 transformed
- Sample size: 100, 250, 500, 1000
- Estimated PAG accuracy, precision, and recall



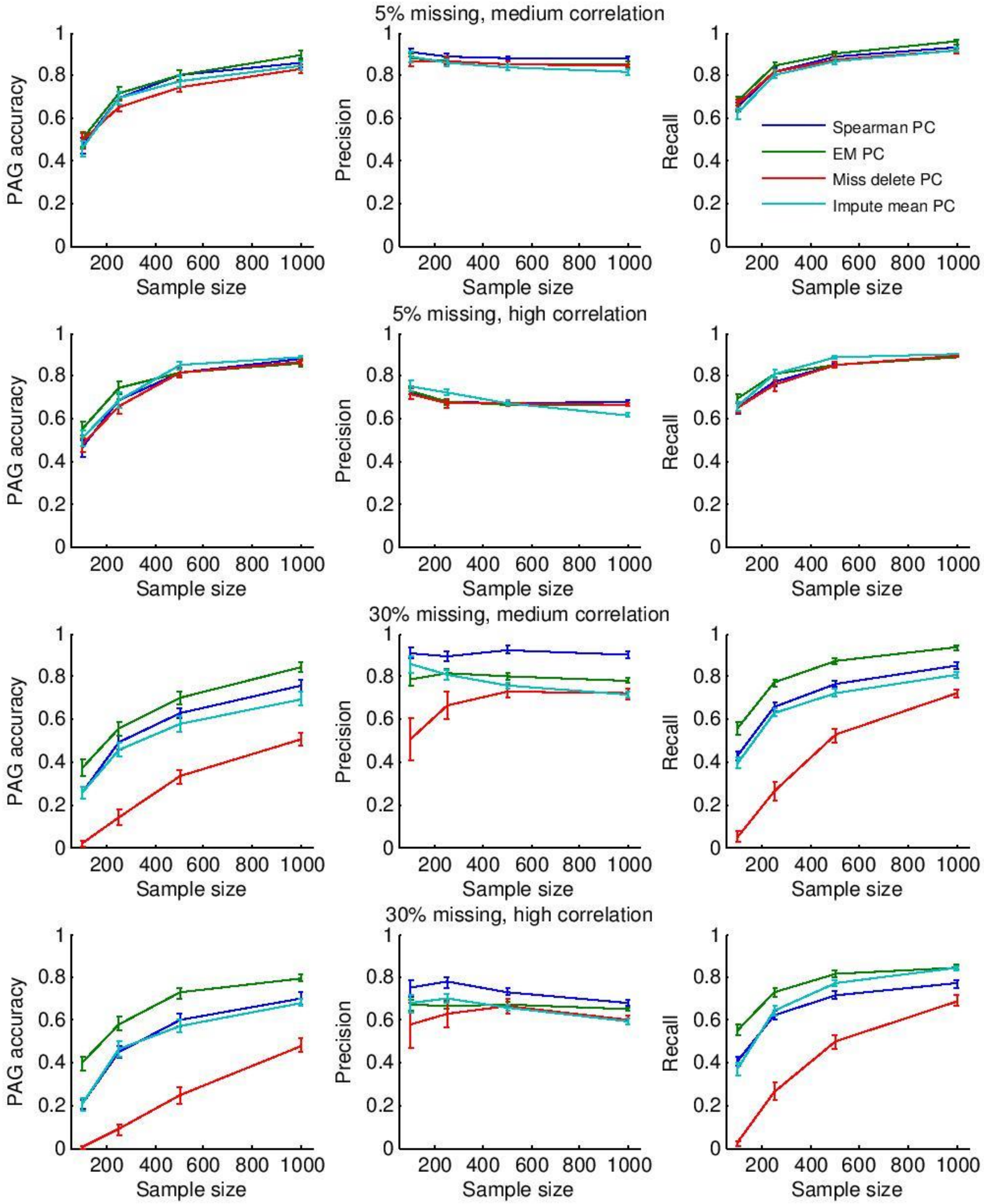
0% missing


0% missing, medium correlation
PAG accuracy
Precision
Recall
Sample size
200 400 600 800 1000
0 0.2 0.4 0.6 0.8 1
Spearman
EM
Spearman adj
0% missing, high correlation

5%, 30% missing BCCD



5%, 30% missing PC



Conclusions

- EM performs better than other methods when there is a significant amount of missing values
- Spearman adjusted leads to unstable matrix and many spurious edges

Real world Data set, ADHD MID task

Type of data:

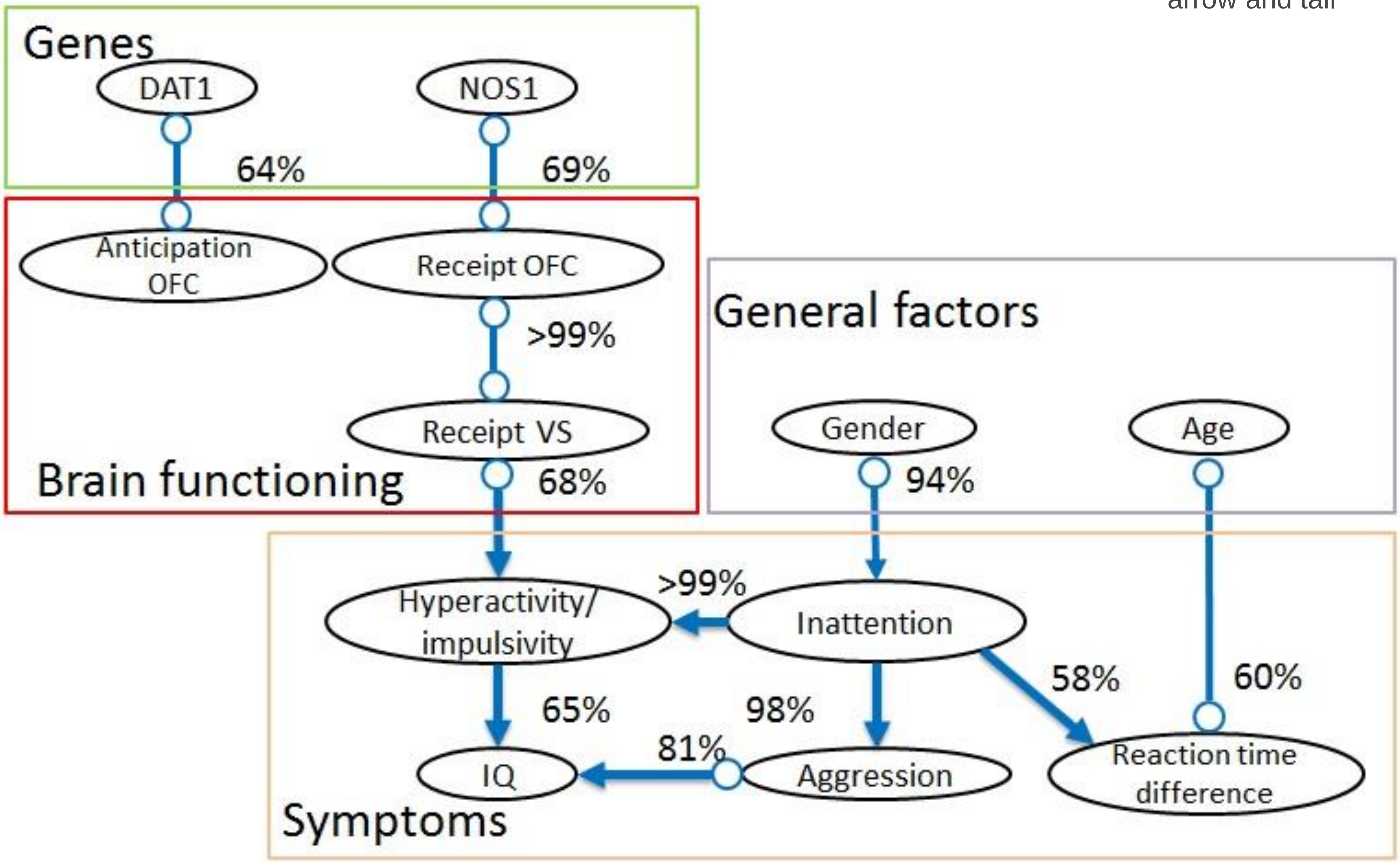
- Genetic information (NOS1, DAT1)
- Brain activation (OFC, VS, anticipation and feedback)
- Behavioral (symptoms, aggression, reaction time, IQ)
- General (age, gender)

Assumptions

- Assumed that missing values are missing at random
- Combined two types of symptoms assessments: by parents and by psychiatrist.
- Incorporated prior knowledge that nothing can cause:
    - Gender
    - Feedback VS is not caused by HI

Real world data ADHD MID task

A → **B**: A causes B
A ↔ **B**: latent common cause
A — **B**: selection bias
O: cannot distinguish between arrow and tail



Real world Data set, ADHD reversal task

Type of data:

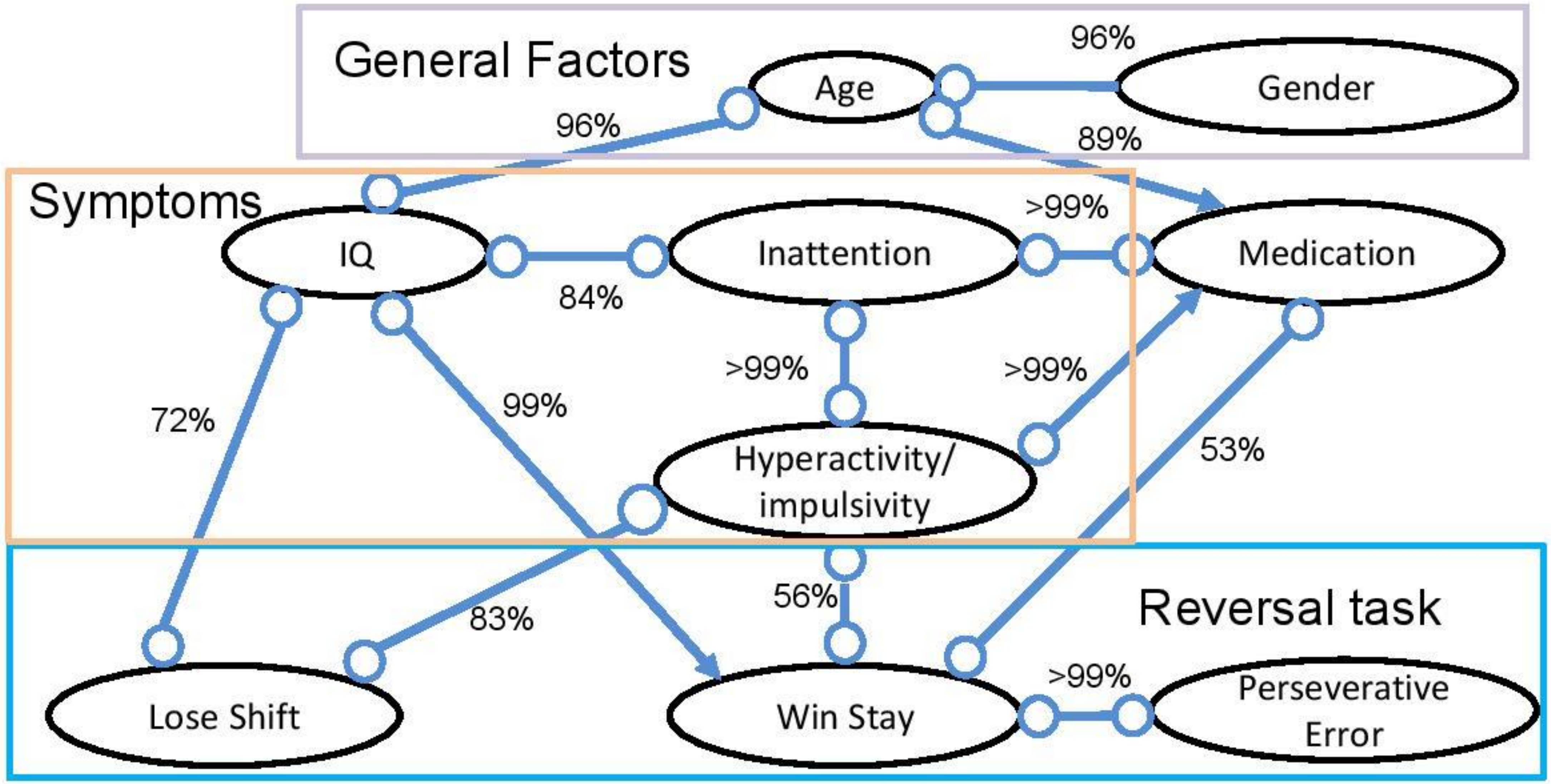
- Experiment related (lose shift, win stay, error)
- Behavioral (symptoms, IQ)
- General (age, gender)

Assumptions

- Assumed that missing values are missing at random
- Incorporated prior knowledge that nothing can cause:
    - Gender

Real world data ADHD reversal task

A → B: A causes B
A ↔ B: latent common cause
A — B: selection bias
O: cannot distinguish between arrow and tail



Conclusions and Future work

- Extension of the BCCD algorithm for mixtures of discrete and continuous variables

- Works well under the assumption of non paranormal data and values MAR

- Further developments:
  - More complex relationships
  - Longitudinal data

Thank you for your attention!