Ancestral Causal Inference

Sara Magliacane, Tom Claassen, Joris M. Mooij

s.magliacane@uva.nl

5^{th} December, 2016

Part I

Introduction

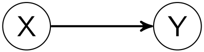
Causal inference: learning causal relations from data

Causal inference: learning causal relations from data

Definition

X causes Y ($X \dashrightarrow Y$) = *intervening upon* (changing) X changes Y

- We can represent causal relations with a causal DAG (hidden vars):



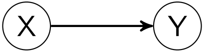
E.g. X = Smoking, Y = Cancer

Causal inference: learning causal relations from data

Definition

X causes Y ($X \dashrightarrow Y$) = *intervening upon* (changing) X changes Y

- We can represent causal relations with a causal DAG (hidden vars):



E.g. X = Smoking, Y = Cancer

- Causal inference = structure learning of the causal DAG

Causal inference: learning causal relations from data

Definition

X causes Y ($X \dashrightarrow Y$) = *intervening upon* (changing) X changes Y

- We can represent causal relations with a causal DAG (hidden vars):

$$X \rightarrow Y$$

E.g. X = Smoking, Y = Cancer

- Causal inference = structure learning of the causal DAG
- Traditionally, causal relations are inferred from interventions.
- Sometimes, interventions are unethical, unfeasible or too expensive

Causal inference from observational and experimental data

Holy Grail of Causal Inference

Learn as much causal structure as possible from observations, integrating background knowledge and experimental data.

Causal inference from observational and experimental data

Holy Grail of Causal Inference

Learn as much causal structure as possible from observations, integrating background knowledge and experimental data.

- **Constraint-based causal discovery**: use statistical independences to express constraints over possible causal models
- **Intuition**: Under certain assumptions, independences in the data correspond with d-separations in a causal DAG

Causal inference from observational and experimental data

Holy Grail of Causal Inference

Learn as much causal structure as possible from observations, integrating background knowledge and experimental data.

- **Constraint-based causal discovery**: use statistical independences to express constraints over possible causal models
- **Intuition**: Under certain assumptions, independences in the data correspond with d-separations in a causal DAG
- **Issues**:
  1. Vulnerability to errors in statistical independence tests
  2. No estimation of confidence in the causal predictions

Causal inference as an optimization problem (e.g. HEJ)

- Weighted list of statistical independence results: $I = \{(i_j, w_j)\}$:
  - E.g. $I = \{\, (Y \perp\!\!\!\perp Z \,|\, X, 0.2),\, (Y \not\perp\!\!\!\perp X, 0.1)\}$

HEJ [Hyttinen et al., 2014]

Causal inference as an optimization problem (e.g. HEJ)

- Weighted list of statistical independence results: $I = \{(i_j, w_j)\}$:
  - E.g. $I = \{\, (Y \perp\!\!\!\perp Z \,|\, X, 0.2),\ (Y \not\!\perp\!\!\!\perp X, 0.1)\}$
- For any possible causal structure C, we define the loss function:

$$\mathcal{L}oss(C, I) := \sum_{(i_j, w_j) \in I:\ i_j \text{ is not satisfied in } C} w_j$$

- "i_j is not satisfied in C" $=$ defined by causal reasoning rules

HEJ [Hyttinen et al., 2014]

Causal inference as an optimization problem (e.g. HEJ)

- Weighted list of statistical independence results: $I = \{(i_j, w_j)\}$:
  - E.g. $I = \{\, (Y \perp\!\!\!\perp Z \,|\, X, 0.2),\ (Y \not\perp\!\!\!\perp X, 0.1)\}$
- For any possible causal structure C, we define the loss function:

$$\mathcal{L}oss(C, I) := \sum_{(i_j, w_j) \in I:\ i_j \text{ is not satisfied in } C} w_j$$

- "i_j is not satisfied in C" = defined by causal reasoning rules
- Causal inference = Find causal structure minimizing loss function

$$C^* = \arg\min_{C \in \mathcal{C}} \mathcal{L}oss(C, I)$$

- **Problem**: Scalability

HEJ [Hyttinen et al., 2014]

Part II

Ancestral Causal Inference

A more coarse grained representation

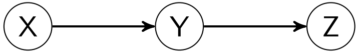
- Can we improve scalability of the most accurate state-of-the-art method (HEJ)?

A more coarse grained representation

- Can we improve scalability of the most accurate state-of-the-art method (HEJ)?

Ancestral Causal Inference: Main Idea

Instead of representing direct causal relations use a more coarse-grained representation of causal information, e.g., an ancestral structure (a set of "indirect" causal relations).

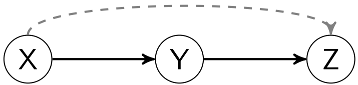


A more coarse grained representation

- Can we improve scalability of the most accurate state-of-the-art method (HEJ)?

Ancestral Causal Inference: Main Idea

Instead of representing direct causal relations use a more coarse-grained representation of causal information, e.g., an ancestral structure (a set of "indirect" causal relations).



- Ancestral structures reduce drastically search space
- For 7 variables: $2.3 \times 10^{15} \rightarrow 6 \times 10^{6}$

Causal inference as an optimization problem (Reprise)

- Weighted list of inputs: $I = \{(i_j, w_j)\}$:
  - E.g. $I = \{\, (Y \perp\!\!\!\perp Z \,|\, X, 0.2),\ (Y \not\perp\!\!\!\perp X, 0.1)\},\ (U \dashrightarrow Z, 0.8)\ \}$
  - Any consistent weighting scheme, e.g. frequentist, Bayesian

Causal inference as an optimization problem (Reprise)

- Weighted list of inputs: $I = \{(i_j, w_j)\}$:
  - E.g. $I = \{\, (Y \perp\!\!\!\perp Z \mid X, 0.2),\ (Y \not\perp\!\!\!\perp X, 0.1)\},\ (U \dashrightarrow Z, 0.8)\ \}$
  - Any consistent weighting scheme, e.g. frequentist, Bayesian
- For any possible ancestral structure C, we define the loss function:

$$\mathcal{L}oss(C, I) := \sum_{(i_j, w_j) \in I:\ i_j \text{ is not satisfied in } C} w_j$$

- Here: "i_j is not satisfied in C" = defined by ancestral reasoning rules

Causal inference as an optimization problem (Reprise)

- Weighted list of inputs: $I = \{(i_j, w_j)\}$:
  - E.g. $I = \{\, (Y \perp\!\!\!\perp Z \,|\, X, 0.2), \, (Y \not\perp\!\!\!\perp X, 0.1)\}, \, (U \dashrightarrow Z, 0.8) \,\}$
  - Any consistent weighting scheme, e.g. frequentist, Bayesian
- For any possible ancestral structure C, we define the loss function:

$$\mathcal{L}oss(C, I) := \sum_{(i_j, w_j) \in I:\ i_j \text{ is not satisfied in } C} w_j$$

- Here: "i_j is not satisfied in C" = defined by ancestral reasoning rules
- Causal inference = Find ancestral structure minimizing loss function

$$C^* = \arg\min_{C \in \mathcal{C}} \mathcal{L}oss(C, I)$$

Ancestral reasoning rules: Example

- ACI rules: 7 ancestral reasoning rules that given (in)dependences constrain possible (non) ancestral relations

Ancestral reasoning rules: Example

- ACI rules: 7 ancestral reasoning rules that given (in)dependences constrain possible (non) ancestral relations

Example

For X, Y, $\boldsymbol{W}$ disjoint (sets of) variables:

$$(X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \not\dashrightarrow \boldsymbol{W}) \implies X \not\dashrightarrow Y$$

- $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ = "X is independent of Y given a set of variables $\boldsymbol{W}$"
- $\wedge$ "and"
- $X \not\dashrightarrow \boldsymbol{W}$ = "X does not cause any variable in the set $\boldsymbol{W}$"
- $\implies$ = "then"
- $X \not\dashrightarrow Y$ = "X does not cause Y"

A method for scoring causal predictions

- Score the confidence in a predicted statement s (e.g. $X \dashrightarrow Y$) as:

$$C(f) = \min_{C \in \mathcal{C}} \mathcal{L}oss(C,\ I + (\neg s, \infty)) - \min_{C \in \mathcal{C}} \mathcal{L}oss(C,\ I + (s,\ \infty))$$

- $\approx$ MAP approximation of the log-odds ratio of s

A method for scoring causal predictions

- Score the confidence in a predicted statement s (e.g. $X \dashrightarrow Y$) as:

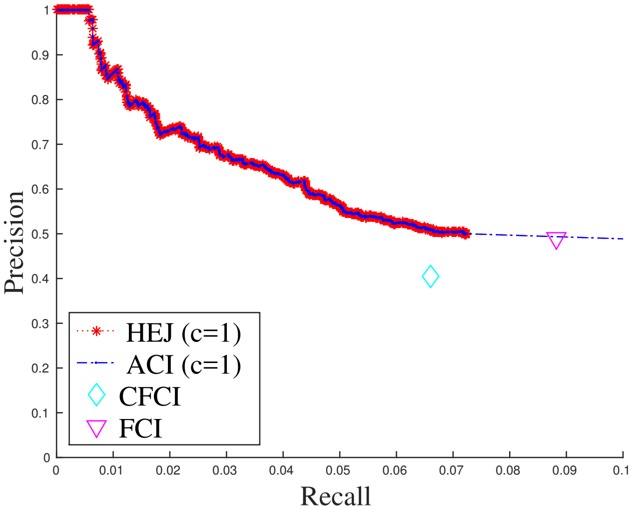
$$C(f) = \min_{C \in \mathcal{C}} \mathcal{L}oss(C,\ I + (\neg s, \infty)) - \min_{C \in \mathcal{C}} \mathcal{L}oss(C,\ I + (s,\ \infty))$$

- $\approx$ MAP approximation of the log-odds ratio of s
- Asymptotically consistent, when consistent input weights
- Can be used with any method that solves an optimization problem

Part III

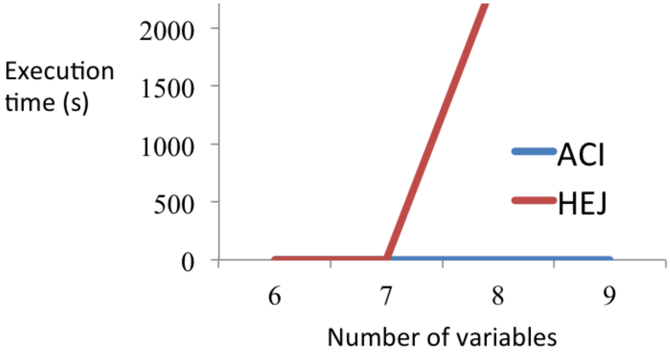
Evaluation

Simulated data accuracy: example Precision Recall curve



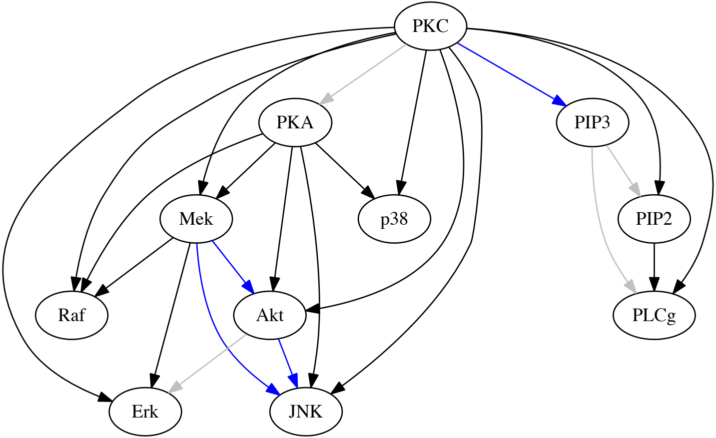
- ACI is as accurate as HEJ + our scoring method

Simulated data execution time



- ACI is orders of magnitude faster than HEJ
- The difference grows exponentially in the number of variables
- HEJ is not feasible for more than 8 variables

Application: Reconstructing a Protein Signalling Network



- Black edges = overlap
- Consistent with score-based method [Mooij and Heskes, 2013]

Conclusion

- Ancestral Causal Discovery (ACI), a causal discovery method as accurate as the state-of-the-art but much more scalable
- A method for scoring causal relations by confidence

- Source code: `http://github.com/caus-am/aci`

- Poster: WIML, 1.30pm - 2.30pm, poster 3
- Poster: NIPS, Tuesday 6pm - 9.30pm, poster 81

- Talk on extensions of ACI at "What If?" NIPS workshop, Saturday

References I

Claassen, T. and Heskes, T. (2011).
A logical characterization of constraint-based causal discovery.
In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 135–144.

Entner, D., Hoyer, P., and Spirtes, P. (2013).
Data-driven covariate selection for nonparametric estimation of causal effects.
In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014).
Constraint-based causal discovery: Conflict resolution with answer set programming.
In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 340–349.

Magliacane, S., Claassen, T., and Mooij, J. M. (2016).
Ancestral causal inference.
arXiv.org preprint, arXiv:1606.07035 [cs.LG].
Accepted for Neural Information Processing Systems 2016.

Mooij, J. M. and Heskes, T. (2013).
Cyclic causal discovery from continuous equilibrium data.
In Nicholson, A. and Smyth, P., editors, *UAI*, pages 431–439. AUAI Press.

Part IV

Backup slides

Infinite weights: Soundness

ACI is sound for predicting ancestral relations:

Theorem

The confidence scores $C(X \dashrightarrow Y)$ are sound for oracle inputs with infinite weights, i.e.:

$$C(X \dashrightarrow Y) = \begin{cases} \infty & \text{if } X \dashrightarrow Y \text{ is identifiable,} \\ -\infty & \text{if } X \not\dashrightarrow Y \text{ is identifiable,} \\ 0 & \text{otherwise.} \end{cases}$$

Finite Weights: Definition and Consistency

We propose two choices for the weights:

- **Frequentist Weights**: $w_j = |\log p_j - \log \alpha|$
  where p_j is the p-value of a statistical test for i_j, and α a threshold.
- **Bayesian Weights**: $w_j = \log p(i_j \mid data) - \log p(\neg i_j \mid data)$.

Under mild assumptions, such weights are consistent, i.e., as sample size $N \to \infty$, for the frequentist weights:

$$\log p^{(N)} - \log \alpha^{(N)} \xrightarrow{P} \begin{cases} -\infty & H_1 \\ +\infty & H_0, \end{cases}$$

when $\alpha^{(N)} \to 0$ at a suitable rate, or for the Bayesian weights:

$$w_N \xrightarrow{P} \begin{cases} -\infty & \text{if } i_j \text{ is true} \\ +\infty & \text{if } i_j \text{ is false.} \end{cases}$$

The probability of a type I and type II errors will then converge to 0.

Finite Weights: Soundness and Consistency

ACI is consistent for predicting ancestral relations:

Theorem

The confidence scores $C(X \dashrightarrow Y)$ are asymptotically consistent, i.e.:

$$C(X \dashrightarrow Y) \xrightarrow{P} \begin{cases} \infty & \text{if } X \dashrightarrow Y \text{ is identifiable,} \\ -\infty & \text{if } X \not\dashrightarrow Y \text{ is identifiable,} \\ 0 & \text{otherwise.} \end{cases}$$

Complete ACI rules

Trivial rules:

1. $X \dashrightarrow Y \wedge Y \dashrightarrow Z \implies X \dashrightarrow Z$ (*transitivity*)
2. $X \dashrightarrow Y \implies Y \not\dashrightarrow X$ (*acyclicity*)

For X, Y, $\boldsymbol{W}$ disjoint (sets of) variables:

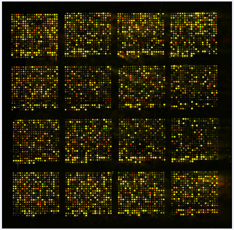
1. $(X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \not\dashrightarrow \boldsymbol{W}) \implies X \not\dashrightarrow Y$
2. $X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [Z] \implies (X \not\perp\!\!\!\perp Z \mid \boldsymbol{W}) \wedge (Z \not\dashrightarrow \{X, Y\} \cup \boldsymbol{W})$
3. $X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [Z] \implies (X \not\perp\!\!\!\perp Z \mid \boldsymbol{W}) \wedge (Z \dashrightarrow \{X, Y\} \cup \boldsymbol{W})$
4. $(X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [Z]) \wedge (X \perp\!\!\!\perp Z \mid \boldsymbol{W} \cup U) \implies (X \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup U)$
5. $(Z \not\perp\!\!\!\perp X \mid \boldsymbol{W}) \wedge (Z \not\perp\!\!\!\perp Y \mid \boldsymbol{W}) \wedge (X \perp\!\!\!\perp Y \mid \boldsymbol{W}) \implies X \not\perp\!\!\!\perp Y \mid \boldsymbol{W} \cup Z$
6. $X \not\dashrightarrow \boldsymbol{W} \wedge X \not\dashrightarrow Z \wedge Z \perp\!\!\!\perp Y \mid \boldsymbol{W} \cup [X] \implies$
   $p\big(Y \mid \mathrm{do}(X)\big) = \int p(Y \mid X, \boldsymbol{W}) p(\boldsymbol{W}) \, d\boldsymbol{W}$

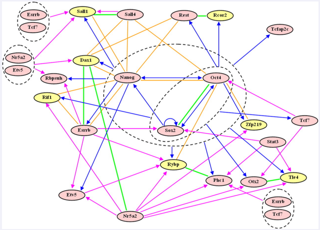
[Claassen and Heskes, 2011], [Entner et al., 2013], [Magliacane et al., 2016]

Applications of causal inference

Example (Genomics)

How to infer gene regulatory networks from micro-array data?





Probabilistic inference vs. causal inference

Traditional statistics, machine learning

- Models the distribution of the data
- Focuses on predicting observations
- Useful e.g. in medical diagnosis: given the symptoms, what is the most likely disease?

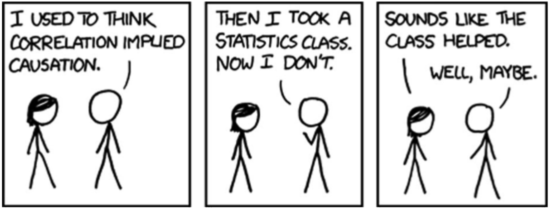
Causal Inference

- Models the mechanism that generates the data
- Also allows to predict results of interventions
- Useful e.g. in medical treatment: if we treat the patient with a drug, will it cure the disease?

Constraint-based causal discovery

Constraint-based: use statistical independences to express constraints over possible causal models.

... but wait, correlation does not imply causation, see XCKD:



True, but it does imply something:

> *If A and B are correlated, A causes B or B causes A or they share a latent common cause. (Hans Reichenbach)*

Idea: Under certain assumptions, independences in the data correspond with d-separations in a causal graph.