Motivation
MLRG Theme
Example
Background
Task
Other approaches
Paper
Setup
Theory
Experiments
Discussion
References

Causal Discovery

UBC MLRG

Betty Shea

11 - Mar - 2020

Agenda

1. Motivation: Causal discovery
   - Within MLRG theme
   - An example
2. Background and Theory
   - Task description
   - Existing approaches
3. Paper: Hoyer et al. (2008)
   - Theoretical results
   - Experimental results

The story so far....

Causal Discovery

Motivation
MLRG Theme
Example

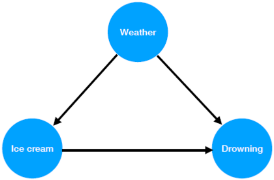
Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

Inference, counterfactual reasoning, confounding factors

- Classical inference techniques e.g. backdoor adjustment (Cathy)
- Counterfactual inference (Ben)
- Instrumental variables (Aaron)
- Inference with VAEs (Wu)



What if we don't have a graph?

We need to find model structure. Causal discovery methods:

- Assume non-Gaussian noise and use independent component analysis (ICA)
- Other approaches
  - Use non-invertibility
  - Markov equivalent DAGs (Sun, Janzing & Schölkopf 2006)
- Today: Use (almost any) non-linearity

One weird trick... statisticians hate this

Causal Discovery

Motivation
MLRG Theme
Example

Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

Main idea:

We want to break the symmetry between observed variables to identify the causal direction.

An example: credit vs stocks

Causal Discovery

Motivation
MLRG Theme
Example

Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

Observation: Credit spreads widen and stocks fall together.

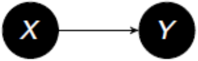
Three competing theories:

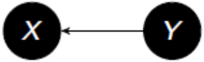
1. Credit spreads widen $\Rightarrow$ stock market selloff
2. Credit spreads widen $\Leftarrow$ stock market selloff
3. Something else causes both

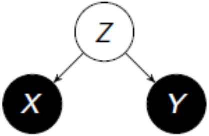
Controlled randomized experiments could be unethical, too expensive or impossible.

Reichenbach's principle of common cause

If two variables X and Y are statistically dependent then either

1. 

2. 

3. 

THE DEATH OF SHERLOCK HOLMES.

Task description

- Every statistical dependence is due to a causal relation.
- May have multiple parents, multiple causal relations.
- Find structure of causality.

General model

Observed variable x_i is a node i in a directed acyclic graph with value

$$x_i := f_i(x_{pa(i)}) + n_i \tag{1}$$

where f_i is an arbitrary function,

$x_{pa(i)}$ is a vector of elements that are parents of x_i,

independent noise variables n_i with arbitrary probability densities p_{n_i}

Special case: Linear model with Gaussian noise

- Observe joint distribution $p(x, y)$
- For linear-Gaussian models, $p(y|x)$ is the same shape as $p(x|y)$
- Hard to distinguish between forward and backward causal directions

Linear model with non-Gaussian noise

Causal Discovery

Motivation
MLRG Theme
Example

Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

If f_i is linear, p_{n_i} is non-Gaussian (Shimizu et al. 2006)

1. Run ICA (PCA using more than covariance information).
2. Factorize $X = AS$. The rows of S contain the independent components. Set $W = A^{-1}$.
3. ICA is not rotation-invariant (with non-Gaussian noise) and so can find factors W.
4. ICA is permutation-invariant and so rows of W are in random order.

Networks with Gaussian priors

If f_i is non-invertible, p_{n_i} is Gaussian (Friedman & Nachman 2000)

- Continuous variable probabilistic networks that are based on Gaussian process priors.
- Interpret learning as assessing the posterior probability of various network structures

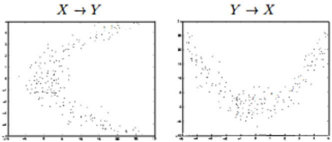


Figure 1: An example of a non-invertible dependence between X and Y. The explanation $X \rightarrow Y$ does not have a functional form, whereas $Y \rightarrow X$ can be explained as a noisy function.

Hoyer et al. (2008)

Causal Discovery

Motivation
MLRG Theme
Example

Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

Nonlinear causal discovery with additive noise models

P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters and B. Schölkopf. (2008) In *Advances in Neural Information Processing Systems 21*: 689-696

- Extends non-invertibility results to any non-linear function
- Noise can follow arbitrary distribution

Structure of the paper

Motivation
MLRG Theme
Example
Background
Task
Other approaches
Paper
Setup
Theory
Experiments
Discussion
References

- Theoretical analysis for 2-variable case. Assumes:
  - strictly positive density functions
  - all functions are thrice differentiable.
- Experimental analysis
  - Simulation with tunable levels of non-linearity
  - Three real world datasets with known causal direction

Theorem 1

Let the joint probability density of x and y be given by

$$p(x, y) = p_n(y - f(x))p_x(x)$$

where p_n, p_x are probability densities on $\mathbb{R}$. If there is a backward model of the same form, i.e.

$$p(x, y) = p_{\tilde{n}}(x - g(y))p_y(y)$$

then denoting $\nu := \log p_n$ and $\xi := \log p_x$, the triple (f, p_x, p_n) must satisfy the following differential equation for all x, y with $\nu''(y - f(x))f'(x) \neq 0$:

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'} \qquad (2)$$

where we have skipped the arguments $y - f(x)$ for ν, x for ξ, and x for f and their derivatives. Moreover, if for a fixed pair (f, ν) there exists $y \in \mathbb{R}$ such that $\nu''(y - f(x))f'(x) \neq 0$ for all but a countable set of points $x \in \mathbb{R}$, the set of all p_x for which p has a backward model is contained in a 3-dimensional affine space.

Theorem 1 - TLDR

Causal Discovery

Motivation
MLRG Theme
Example

Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

- The space of all possible log-marginals ξ is infinite dimensional.
- Fixing ξ, ξ' and ξ" at some arbitrary point x_0 will completely determine ξ
- ξ has a 3-dimensional space of solutions.
- Therefore, forward model cannot be inverted and true model (causality direction) is identifiable.

Corollary 1

Causal Discovery

Motivation
MLRG Theme
Example

Background
Task
Other approaches

Paper
Setup
Theory
Experiments
Discussion

References

Assume that $\nu''' = \xi''' = 0$ everywhere.

If a backward model exists, then f is linear.

Experimental goal

Empirical tests try to distinguish these four scenarios:

- observable variables are mutually independent (1)
- observable variables are dependent and
  - there are conflicting causal directions (2)
  - there are no causal direction (3)
  - there is only one causal direction (4)

Experimental procedure

For each DAG G_i (forward and backward)

1. non-linear regression of each variable on its parents to learn $\hat{f}$
2. test for independence of residual $\hat{n} = y - \hat{f}(x)$ with x
3. Reject G_i if any independence test fails. Accept otherwise.

Feasible for only very small networks.

Suffers from the problem of multiple hypothesis testing

Simulations

- Data simulated using $y = x + bx^3 + n$
  - x and n are sampled from Gaussian distribution
  - x and n raised to the power q while keeping original sign
- b controls strength of non-linearity.
- q controls how close to Gaussian the noise is
- Hypothesis testing with 2% significance level
- For each combination of b and q, repeat experiment $100\times$

Simulation results

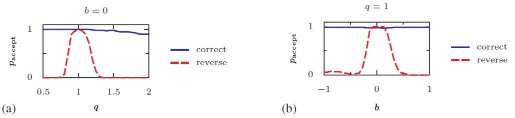


Figure 2: Results of simulations (see main text for details): **(a)** The proportion of times the forward and the reverse model were accepted, p_{accept}, as a function of the non-Gaussianity parameter q (for $b = 0$), and **(b)** as a function of the nonlinearity parameter b (for $q = 1$).

Model able to infer the correct causal direction either when

- distributions are sufficiently non-Gaussian
- distributions are sufficiently non-linear

Motivation
MLRG Theme
Example
Background
Task
Other approaches
Paper
Setup
Theory
Experiments
Discussion
References

Real-world data

Datasets:

- Old Faithful: duration of an eruption and the time interval between subsequent eruptions
- Abalone: number of rings in the shell and length of the shell
- Altitude-temperature: altitude above sea level and local yearly average outdoor temperature

Real-world data results

Method picks:

- forward model "current duration causes next interval length" and not backward model "next interval length causes current duration"
- age causes length of shell and not length of shell causes age
- altitude causes temperature over vice versa

Real-world data results

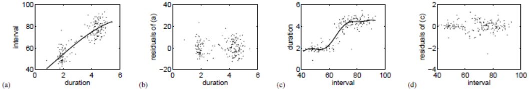


Figure 3: The Old Faithful Geyser data: (a) forward fit corresponding to "current duration causes next interval length"; (b) residuals for forward fit; (c) backward fit corresponding to "next interval length causes current duration"; (d) residuals for backward fit.

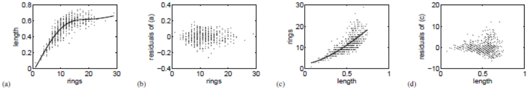


Figure 4: Abalone data: (a) forward fit corresponding to "age (rings) causes length"; (b) residuals for forward fit; (c) backward fit corresponding to "length causes age (rings)"; (d) residuals for backward fit.

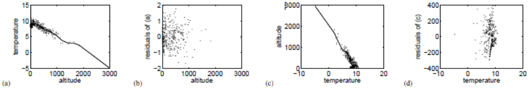


Figure 5: Altitude–temperature data. (a) forward fit corresponding to "altitude causes temperature"; (b) residuals for forward fit; (c) backward fit corresponding to "temperature causes altitude"; (d) residuals for backward fit.

Some questions

- Does real world data fit the criteria of non-linear f or non-Gaussian residuals?
- How do we pick value for acceptance of null hypothesis?
- Is the thrice differentiable requirement reasonable?
- How realistic is it to assume that noise is independent?

References

Chaves, R., Luft, I., Maciel, T.O., Gross, D., Janzing, D. & Schökopf, B. (2014) Inferring latent structures via information inequalities. *UAI*

Friedman, N. & Nachman, I. (2000) Gaussian process networks. In *Proc. of the 16th Annual Conference on Uncertainty in Artificial Intelligence*: 211-219

Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J. & Schölkopf, B. (2008) Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*: 689-696.

Janzing, D. (2019) Non-statistical notions of independence in causal discovery. https://www.groups.ma.tum.de/fileadmin/w00ccg/statistics/veranstaltungen/Graphical_Models__Conditional_Independence_and_Algebraic_Structures/Janzing__Dominik.pdf

References

Peters, J., Janzing, D. & Schölkopf, B. (2017) Elements of Causal Inference. *Available through Open Access*: https://mitpress.mit.edu/books/elements-causal-inference

Shimizu, S., Hoyer, P.O., Hyvärinen, A. & Kerminen, A.J. (2006) A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, **7**: 2003-2030.

Steudel, B., Janzing, D. & Schölkopf. (2010) Causal Markov condition for submodular information measures. *COLT 2010*: 464-476

Sun, X., Janzing, D. & Schölkopf, B. (2006) Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th Int. Symp. Art. Int. and Math.*


That's all Folks!