Distinguishing between Cause and Effect: Estimation of Causal Graphs with two Variables

Jonas Peters
ETH Zürich

Tutorial

NIPS 2013 Workshop on Causality
9th December 2013
Eating chocolate produces Nobel prize winners, says study

By Oliver Nieburg, 11-Oct-2012

Related tags: noble prize, noble laureate, Einstein, Marie Curie, chocolate, brain, Switzerland, Sweden, candy

You don’t have to be a genius to like chocolate, but geniuses are more likely to eat lots of chocolate, at least according to a new paper published in the August New England Journal of Medicine. Franz Messerli reports a highly
Problem: Given $P(X, Y)$, can we infer whether

$$X \rightarrow Y \quad \text{or} \quad Y \rightarrow X?$$
Problem: Given $P(X, Y)$, can we infer whether

\[ X \rightarrow Y \quad \text{or} \quad Y \rightarrow X ? \]

Difficulty: So much symmetry:

\[ P(X) \cdot P(Y \mid X) = P(X, Y) = P(X \mid Y) \cdot P(Y) \]

We need assumptions!! (e.g. Markov and faithfulness do not suffice.)
**Problem:** Given $P(X, Y)$, can we infer whether

$$X \rightarrow Y \quad \text{or} \quad Y \rightarrow X?$$

**Difficulty:** So much symmetry:

$$P(X) \cdot P(Y | X) = P(X, Y) = P(X | Y) \cdot P(Y)$$

We need assumptions!! (e.g. Markov and faithfulness do not suffice.)

**Surprise** (for some assumptions):

$$2 \text{ variables} \Rightarrow p \text{ variables}$$

Idea No. 1: Linear Non-Gaussian Additive Models (LiNGAM)

Structural assumptions like additive non-Gaussian noise models break the symmetry:

\[ Y = \beta X + N_Y \quad N_Y \perp \! \! \! \! \perp X, \]

with \( N_Y \) non-Gaussian.
Consider a distribution corresponding to

\[ Y = \beta X + N_Y \]

- \( N_Y \perp X \)
- \( N_Y \) non-Gaussian
Consider a distribution corresponding to

\[ Y = \beta X + N_Y \]

- \( N_Y \perp \!\!\!\!\perp X \)
- \( N_Y \) non-Gaussian

Then there is no

\[ X = \phi Y + N_X \]

- \( N_X \perp \!\!\!\!\perp Y \)
- \( N_X \) non-Gaussian

Idea No. 2: Additive noise models

**Nonlinear functions** are also fine!

\[ Y = f(X) + N_Y \quad N_Y \perp X \]

Asymmetry No. 2

Consider a distribution corresponding to

\[ Y = f(X) + N_Y \]

with \( N_Y \perp\!\!\!\!\!\!\!\perp X \)
Consider a distribution corresponding to
\[ Y = f(X) + N_Y \]
with \( N_Y \perp X \)

Then for “most combinations” \((f, P(X), P(N_Y))\) there is no
\[ X = g(Y) + M_X \]
with \( M_X \perp Y \)
\[ Y = f(X) + N_Y, \quad N_Y \perp X \]
\[ Y = f(X) + N_Y, \quad N_Y \perp \perp X \]
\[ X = g(Y) + N_X, \quad N_X \perp \perp Y \]
\[ X = g(Y) + N_X, \quad N_X \perp Y \]
Idea No. 3: Gaussian Process Inference (GPI)

We can always write

\[ Y = f(X, N_Y), \quad N_Y \perp\!\!\!\!\perp X \]

and

\[ X = g(Y, N_X), \quad N_X \perp\!\!\!\!\perp Y \]

Which model is more “complex”? Use Bayesian model comparison.

J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, B. Schölkopf:  
*Probabilistic latent variable models for distinguishing between cause and effect*, NIPS 2010

Asymmetry No. 3

1. Fix the noise distribution to be $\mathcal{N}(0, 1)$.
2. Put prior $p(\theta_X)$ on input distribution $p(x \mid \theta_X)$ ($\leadsto$ complexity of $X$).
3. Put prior $p(\theta_f)$ on the functions $p(f \mid \theta_f)$ ($\leadsto$ complexity of $f$).
1. Fix the noise distribution to be $\mathcal{N}(0,1)$.

2. Put prior $p(\theta_X)$ on input distribution $p(x \mid \theta_X)$ ($\sim$ complexity of $X$).

3. Put prior $p(\theta_f)$ on the functions $p(f \mid \theta_f)$ ($\sim$ complexity of $f$).

4. Approximate marginal likelihood for $X \to Y$

$$p(x, y) = p(x) \cdot p(y \mid x)$$

$$= \int p(x \mid \theta_X)p(\theta_X) \, d\theta_X$$

$$\cdot \int \delta(y - f(x, e))p(e)p(f) \, de \, df$$

5. Approximate marginal likelihood for $Y \to X$.

6. Compare.

J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, B. Schölkopf:

*Probabilistic latent variable models for distinguishing between cause and effect*, NIPS 2010
Idea No. 4: Information Geometric Causal Inference (IGCI)

Assume a deterministic relationship

\[ Y = f(X) \]

and that \( f \) and \( P(X) \) are “independent”.

D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, B. Schölkopf:

*Information-geometric approach to inferring causal directions*, Artificial Intelligence 2012
Idea No. 4: Information Geometric Causal Inference (IGCI)

Assume a deterministic relationship

\[ Y = f(X) \]

and that \( f \) and \( P(X) \) are “independent”.

D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, B. Schölkopf:

*Information-geometric approach to inferring causal directions*, Artificial Intelligence 2012
Consider $Y = f(X)$ with $id \neq f : [0, 1] \rightarrow [0, 1]$ invertible and $X = g(Y)$. If

$$\text{“cov”} (\log f', p_X) = \int \log(f'(x)) p_X(x) \, dx - \int \log f'(x) \, dx = 0$$

then

$$\text{“cov”} (\log g', p_Y) = \int \log(g'(y)) p_Y(y) \, dy - \int \log g'(y) \, dy > 0$$

D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, B. Schölkopf:

*Information-geometric approach to inferring causal directions*, Artificial Intelligence 2012
Open Questions 1: Quantifying Identifiability

\[ \{Q: Y \rightarrow X\} \]

\[ Q^* = \arg\inf KL(P||Q) \]

\[ \{P: X \rightarrow Y\} \]
Proposition

Assume $P(X, Y)$ is generated by

$$Y = f(X) + N_Y$$

with independent $X$ and $N_Y$.

Then

$$\inf_{Q \in \{Q: Y \rightarrow X\}} \text{KL}(P \parallel Q) = ?$$

- first steps to understand the geometry
- gives us finite sample guarantees
What happens if assumptions are violated? E.g., in case of confounding?

Can we still infer $X \rightarrow Y$? How useful is this?
Conclusions

In theory, we can brake asymmetry between cause and effect.

- restricted structural equation models:
  - linear functions, additive non-Gaussian noise
  - nonlinear functions, additive noise
- complexity measures on functions and distributions
- “independence” between function and input distribution
Conclusions

In theory, we can brake asymmetry between cause and effect.

- restricted structural equation models:
  - linear functions, additive non-Gaussian noise
  - nonlinear functions, additive noise
- complexity measures on functions and distributions
- “independence” between function and input distribution
- ... principles behind new methods from challenge?
Conclusions

In theory, we can brake asymmetry between cause and effect.

- restricted structural equation models:
  - linear functions, additive non-Gaussian noise
  - nonlinear functions, additive noise
- complexity measures on functions and distributions
- “independence” between function and input distribution
- ... principles behind new methods from challenge?

Causal inference problem of climate change is solved! Fight the cause!

Don’t fly! (Zurich-SFO 5.4t CO₂)! Compensate!
It turns out that if $X \to Y$

$$\int \log |f'(x)| p(x) \, dx < \int \log |g'(y)| p(y) \, dy$$

Estimator:

$$\hat{C}_{X \to Y} := \frac{1}{m} \sum_{j=1}^{m} \log \left| \frac{y_{j+1} - y_j}{x_{j+1} - x_j} \right| \approx \int \log |f'(x)| p(x) \, dx$$

Infer $X \to Y$ if

$$\hat{C}_{X \to Y} < \hat{C}_{Y \to X}$$
\[ Y = \beta X + N_Y, \quad N_Y \perp \!\!\!\!\!\!\!\!\!\!\perp X, \quad N_Y \text{ non-Gaussian} \]
\[ Y = \beta X + N_Y, \quad N_Y \perp X, \quad N_Y \text{ non-Gaussian} \]
\[ X = \phi Y + N_X, \quad N_X \perp Y, \quad N_X \text{ non-Gaussian} \]
\[ X = \phi Y + N_X, \quad N_X \perp \perp Y, \quad N_X \text{ non-Gaussian} \]
Does $X$ cause $Y$ or vice versa?

Real Data
Does $X$ cause $Y$ or vice versa?

Real Data
Does $X$ cause $Y$ or vice versa?

No (not enough) data for chocolate
Does $X$ cause $Y$ or vice versa?

No (not enough) data for chocolate

... but we have data for coffee!
Correlation: 0.698, $p$-value: $< 2.2 \cdot 10^{-16}$.
Does X cause Y or vice versa?

Correlation: 0.698, p-value: $< 2.2 \cdot 10^{-16}$.

Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of 0).
Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of 0).

$\Rightarrow$ Model class too small? Causally insufficient?
The linear Gaussian case

\[ Y = \beta X + N_Y \]

with independent
\[ X \sim \mathcal{N}(0, \sigma_X^2) \quad \text{and} \quad N \sim \mathcal{N}(0, \sigma_N^2) \]

Then there is a linear SEM with
\[ X = \alpha Y + M_X \]

How can we find \( \alpha \) and \( M_X \)?

Jonas Peters (ETH Zurich)
Distinguishing between Cause and Effect
9th December 2013