Causal Inference from Noise

26 June 2017
Nevin Climenhaga, University of Notre Dame
Lane DesAutels, Missouri Western State University
Grant Ramsey, KU Leuven

Section 1: Randomized Controlled Trials

- Suppose we are interested in whether diet soda causes type 2 diabetes. The historical data reveal a correlation between the consumption of diet soft drinks and type-2 diabetes.
    - Problems in inferring from this that diet soda causes type 2 diabetes:
        - The direction of causation might go the other way: type-2 diabetes → increases thirst → increases diet soft drink consumption.
        - Causal confounders: e.g., diet soda consumers tend to be lower in economic status → consume more calorie dense processed foods → type-2 diabetes.
- Statistical Orthodoxy: we need to conduct a *Randomized Controlled Trial* (RCT).
    - We need an *experimental* group provisioned with and assigned to drink diet sodas, and a *control* group who does not drink diet soda.
    - Fisher's Guarantee: "The full procedure of randomization [is the method] by which the validity of the test of significance may be *guaranteed* against corruption by the causes of disturbance which have not been eliminated" (1947, p. 19; italics added).
- Reasons to be wary of RCTs:
    - Difficult to *isolate* the cause in such a way that the experimental and control groups have only a single difference.
    - *Economic* and *ethical* problems for researchers.
    - *Randomization is no guarantee* for eliminating the confounding causal factor.
        - Results of randomization might be just as loaded.
        - Re-randomizing re-opens the door to selection bias, and is only possible with antecedent knowledge of possible confounders.
        - The only recourse would be to reach the limiting average by re-randomizing indefinitely. But this is a practical impossibility.
- Upshot: Without knowledge of what these confounders are, RCTs may not have a big advantage over historical trials.

Section 2: Noise Inference Methods (NIMs)

- Case 1: Max the mathematician has placed you in a room with two computers, X and Y. Max controls the number displayed on one of the computers directly. This computer sends a signal to the other computer, so that either $y = x^3$, or $x = \sqrt[3]{y}$. However, there is an occasional glitch leading the output computer to display a number 1 greater than the value it is supposed to display.
    - Observation t_3 is inconsistent with X←Y, but consistent with X→Y.
- Case 2: Same as Case 1, except that there is a third possibility: the values on both Y and X are determined by a third hidden computer Z, according to the functions $x = z^3$ and $y = z^9$.
    - Observation t_3 is inconsistent with X←Y, and t_4 is inconsistent with X→Y. Both are consistent, however, with X←Z→Y.

	X	***Y***
t_1	.5	.125
t_2	-11	-1331
t_3	4	65

Your observations in Case 1

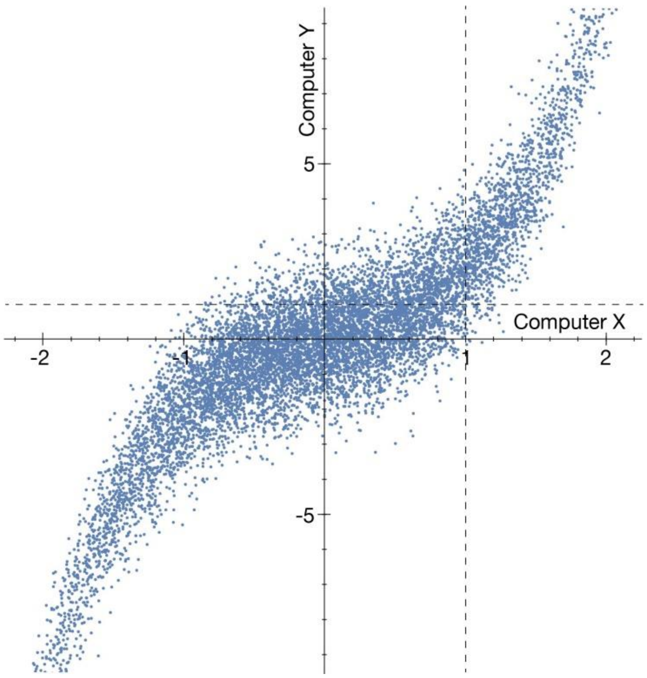
	X	***Y***
t_1	.5	.125
t_2	-11	-1331
t_3	4	65
t_4	3	8

Your observations in Case 2

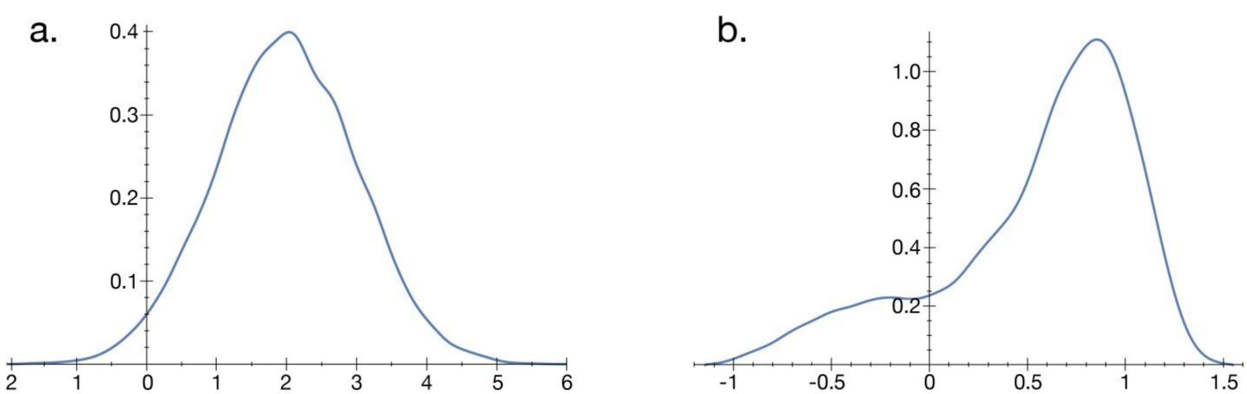
- Case 3: Same as Case 1, except that now (a) Max does not tell you what the mathematical function leading from X to Y or Y to X is, and (b) the glitch is now Gaussian noise – that is, the value on the output computer differs from the true value by a particular amount with the probability represented by a normal Gaussian distribution (bell curve).
    - In general, you find that as you collect more data, when X = x, the distribution of y-values looks roughly like a bell curve centered on $x + x^3$. However, in general, the shape approached by your distribution of x-values when Y = y is not a bell curve.
    - If Y were causing X according to some function $x = f(y)$ + noise, then the frequency with which X has a particular value when Y has a particular value should be a bell curve centered on $f(y)$, since the glitch leads the output to deviate from its true value with Gaussian noise. However, you observe the converse:

when Y has a particular value, the frequency with which X takes on various values is Gaussian. This is consistent with X causing Y according to the function $y = g(x)$ + noise. Moreover, in general the most frequent Y-value when X = x is $x + x^3$, suggesting that $g(x) = x + x^3$.

- The hypothesis that X→Y makes a prediction which you have found to be true, while the hypothesis that X←Y makes a prediction you have found to be false. So in this case, you can again conclude, solely from correlational data, that X→Y.



Your observations in Case 3



a. *The observed distribution of 10,000 y-values when X = 1.*
b. *The observed distribution of 10,000 x-values when Y = 1.*

Section 3: Practical Implications

- Ways in which real world examples differ from above thought experiments: causal relationships more complicated, different causal hypotheses not mutually exclusive, common causes ubiquitous, nature of causal functions and noise (e.g., whether noise is Gaussian) often uncertain.
- In practice, the best way to determine effectiveness of NIMs is to test them on real world and simulated data sets – e.g., Mooij et al. (2016: 58) test NIMs on a data set consisting of "measurements of altitude and mean annual temperature of more than 300 weather stations in Germany." In this case we know that altitude→temperature, rather than altitude←temperature. Mooij et al. find that the most effective NIMs accurately get this result between 63% and 85% of the time.

Section 4: Theoretical Implications

- Various conditions proposed as necessary for causal inferences, such as RCTs or mechanistic knowledge, are not always necessary.
- Noise is standardly defined as "data received but unwanted" (Floridi 2016). But in some situations, we can make causal inferences from the data only if it is noisy. Noise is not unwanted, nor do we choose not to explain it.
    - Possible alternative definition of noise as deviation from "ideal" causation (signaling) from a cause (source) to an effect (receiver).
- While sometimes practically unavoidable, it can be dangerous to throw away data, as RCTs do.

References

Fisher, R. A. 1947. *The Design of Experiments.* 4th edition. Edinburgh: Oliver and Boyd.

Floridi, Luciano, "Semantic Conceptions of Information," *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2016/entries/information-semantic/>.

Mooij, Joris M., Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. "Distinguishing cause from effect using observational data: methods and benchmarks." *Journal of Machine Learning Research* 17(32): 1-102.