# Inferring Causal Directions in Two Variables

Lohit Jain[1]
Balram Meena[1]
Advisor: Dr. Amitabha Mukerjee[1]
{lohit,balram,amit}@iitk.ac.in
[1] *Dept. Of Computer Science and Engineering*

March 5, 2014

## 1. Introduction

In many aspects of modern everyday life, attributing causes to effects has become very important and pervasive task. Prevalent method for this causality inference is to perform costly experiments, but these collection data are readily available. Inferring causality from this data would result in considerably less time and effort spent in this.
For example, consider lung cancer. The objective is to find a factor, like "*Smoking*", that might be the cause of the effect of "*Lung Cancer*".
But a dependency between A and B does not necessarily imply causality inference. They could both be effects of common cause.

## 2. The Problem

Given the joint distribution of two variables A, B, find the causality.
In other words, find whether A => B or B => A or neither.

For example, given a distribution of *Altitude of cities* and *Average Temperature*. We need to infer that *Altitude* is the cause and *Temperature* is the effect.

This problem was part of ***Causality Challenge*** conducted in 2013.

## 3. Proposed Solution

The Additive noise models [2], linear acyclic models and post linear models pose some additional restriction on the functional dependence between the two given random variables. But in practice, a limited model class may lead to wrong conclusion about the causal direction.

The model which we are going to use does not have added assumptions. It treats the noise as latent variable, which contains the effects of all the unobserved data. We assume it to be independent of all other variables.
To summarize, if Y is effect of X, we model,

$$Y = f(X, E)$$

The broad idea, I propose to use as given by [3], is to define non-parametric priors on the causal mechanisms and input distributions and then infer the causality using standard Bayesian model selection. To decide upon causal direction, or even the existence of causality, we will compare evidence of the models corresponding to each hypothesis X=>Y and Y=>X.

## 4. Data Set and Evaluation

The datasets will be from **Causality Challenge**.
The data sets consists of mix of artificially generated data and pairs of real variables with known causal relationships from domains as diverse as chemistry, climatology, ecology, economy, engineering, epidemiology, genomics, medicine, physics. and sociology.

The evaluation scheme will be based on the scheme used by **Causality Challenge 2013**, namely AUC (Area under the ROC Curve).

## References


[1] Hoyer, Patrik O., et al. "Nonlinear causal discovery with additive noise models." NIPS. Vol. 21. 2008.

[2] Peters, Jonas, Dominik Janzing, and Bernhard Scholkopf. "Causal inference on discrete data using additive noise models." Pattern Analysis and Machine Intelligence, IEEE Transactions on 33.12 (2011): 2436-2450.

[3] Mooij, Joris M., et al. "Probabilistic latent variable models for distinguishing between cause and effect." NIPS. 2010.

[4] Daniusis, Povilas, et al. "Inferring deterministic causal relations." arXiv preprint arXiv:1203.3475 (2012).

[5] Janzing, Dominik, et al. "Information-geometric approach to inferring causal directions." Artificial Intelligence 182 (2012): 1-31.