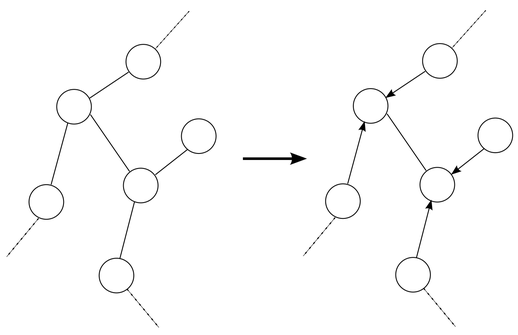
Inferring causal relationships using information-theoretic measures

Catharina Olsen, Patrick E. Meyer and Gianluca Bontempi

December 2009

Machine Learning Group
Université Libre de Bruxelles
Belgium

Introduction



Problem Given microarray datasets, is it possible to infer a graphical representation of the underlying data generating process? State-of-the-art causal inference methods are non-adaptable to situations where hundreds and possibly thousands of variables are involved as in microarray datasets. Starting with the correct adjacency network as a basis, we will present a method which is capable to orient the edges when dealing with a high-variable, low-sample setup exhibited by microarray data.

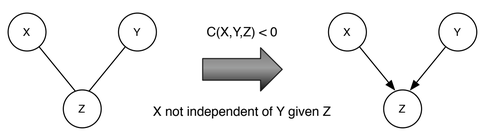
Preliminaries

If three nodes X, Y and Z are linked as $X \rightarrow Z \leftarrow Y$, it is known as **v-structure**, and the node Z is then called a **collider**.
State-of-the-art algorithms use conditional independence tests to deduce the orientation of the arcs [2]. In these algorithms, $X \not\perp\!\!\!\perp Y|Z$ implies either that Z is a collider or that there is a larger conditioning set which renders the variables independent.

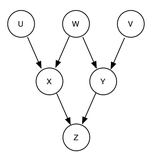
Definition 1 *Given three random variables X, Y and Z, the **interaction interaction** between these variables is defined as*

$$C(X,Y,Z) = I(X,Y) - I(X,Y|Z).$$

Once it is known that $X - Z - Y$, then a negative interaction information implies that Z is a collider.



In more complicated cases, the interaction information provides additional information. If the mutual information is larger than zero, our method may still be able to infer colliders.



Here the independence tests (conditioning on one variable) will result in $W \not\perp\!\!\!\perp Z|X$, $W \not\perp\!\!\!\perp Z|Y$ and $X \not\perp\!\!\!\perp Y|Z$. Thus, there is a contradiction in which node should be a collider. The interaction information however gives a degree of dependence and hence is able to infer the collider.

Method

Given a dataset, our causal inference method advances in three steps. (In the experimental section, we will focus on Step 3.)

1. Inference of the undirected network using a method able to cope with the usually high number of variables (e.g. ARACNE, CLR or MRNET).
2. Estimate the interaction information for all possible v-structures.
3. **Use the interaction information criterium to orient the v-structures.**

Hyper-parameters Possible parameters which influence the performance of the orientation phase.

1. θ is threshold below which triplets will not be oriented.
2. Sort order: with which triplet should the orienting start.
3. Estimator of interaction information (we use the Schürmann-Grassberger estimator applied to the discretized data).

Order of orientation The interaction information can be estimated for every (unshielded) triplet of variables in the adjacency network. In the experimental section, we will show that the performance depends largely on the order of orientation. Three different methods will be tested here to show the impact of the ordering.

1. Random selection of triplets (RAND).
2. Orient in decreasing order of (neg.) interaction information values (MAX).
3. Orient in decreasing order of average (neg.) interaction information (AVG).

Assumptions In order to infer the causal relationships as described, the following assumptions ensure that dependencies/independencies can be used to deduce the causal influences between the variables: (a) causal sufficiency, (b) causal Markov and (c) faithfulness assumptions [1, 3].

Experimental setup

Data We use microarray synthetic data from different sources: (a) Syntren data generator, (b) GeneNetWeaver (GNW) data generator (DREAM challenge) and (c) the LUCAP dataset stemming from the NIPS 2008 challenge.

Dataset	# variables	# samples	# samples/run	origin
LUCAP	144	2000	200	causal Bayesian network
Syntren	300	800	100	EColi
GNW	3000	3000	300	Yeast

For each of the ten runs, a subset of samples was drawn at random from the full dataset. Different values for the parameter θ were tested.

Results

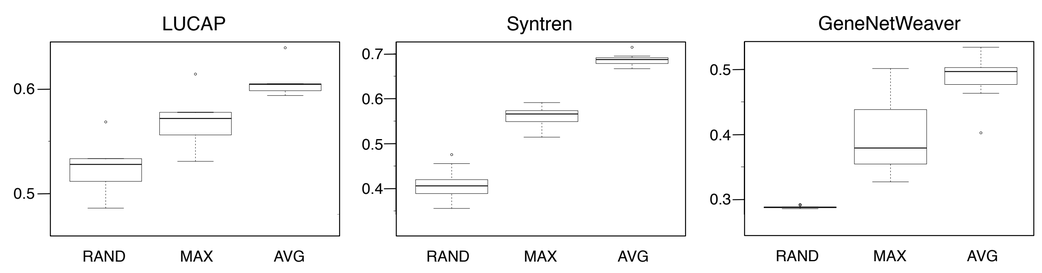
As a performance measure the $F_{0.5}$-score is used

$$F_\beta = \frac{(1+\beta^2)\cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}.$$

Note $F_\beta \in [0,1]$, where 1 is obtained for the correct network.
In the table, the maximum $F_{0.5}$-score (averaged over ten runs) is displayed, the threshold was chosen to be in the interval $[-0.1, 0.1]$.

Dataset/Method	RAND	MAX	AVG
LUCAP	0.5255	0.5701	0.6082
Syntren	0.4081	0.5595	0.6864
GeneNetWeaver	0.0767	0.3048	0.4774

Boxplots for the three datasets: LUCAP, Syntren, GeneNetWeaver. Each graph contains the boxplots for the random, maximum and average selection.



- Taking a node's contect is important when orienting the arcs (AVG method works the best on all datasets).
- The performance depends on the chosen threshold.
- The higher the number of variables, the higher the impact of the ordering.

References

[1] J. Pearl. *Causality : Models, Reasoning, and Inference*. Cambridge University Press, March 2000.

[2] B. Shipley. *Cause and Correlation in Biology : A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, August 2002.

[3] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, MA, USA, second edition, January 2001.