A Statistic Criterion for Reducing Indeterminacy in Linear Causal Modeling

Gianluca Bontempi
Machine Learning Group, Computer Science Department, Faculty of Sciences, Université Libre de Bruxelles, Brussels, Belgium

Keywords: Graphical Models, Causal Inference, Feature Selection.

Abstract: Inferring causal relationships from observational data is still an open challenge in machine learning. State-of-the-art approaches often rely on constraint-based algorithms which detect v-structures in triplets of nodes in order to orient arcs. These algorithms are destined to fail when confronted with completely connected triplets. This paper proposes a criterion to deal with arc orientation also in presence of completely linearly connected triplets. This criterion is then used in a Relevance-Causal (RC) algorithm, which combines the original causal criterion with a relevance measure, to infer causal dependencies from observational data. A set of simulated experiments on the inference of the causal structure of linear networks shows the effectiveness of the proposed approach.

1 INTRODUCTION

One of the most difficult aspect of causal inference from observational data is the indeterminacy of causal structures, due to the existence of dependency structures implying different causal directions but which are indistinguishable in terms of statistical likelihood or fit indexes. For instance it is well known that the detection of causal directionality requires strong assumptions (e.g. nonlinearity, high dimensional observations) in a bivariate (i.e., single cause single effect) context (Janzing et al., 2010; Janzing et al., 2011). This is the reason why existing techniques address triplet configurations to reconstruct the directionality of the causal relationships. Well-known examples are the algorithms which infer causal structures in Bayesian networks by searching for unshielded colliders (Spirtes et al., 2000), i.e. patterns where two variables are both direct causes of a third one, without being each a direct cause of the other. Under assumptions of Causal Markov Condition and Faithfulness, this structure is statistically distinguishable and so-called constraint based algorithms (notably the PC and the SGS algorithms) rely on conditional independence tests to orient at least partially a graph (Koller and Friedman, 2009).

Other research works take advantage of conditional independence and propose information theoretic methods for network inference and feature selection (Brown, 2009; Watkinson et al., 2009; Bontempi and Meyer, 2010; Bontempi et al., 2011). In particular these works use the notion of feature interaction, a three-way mutual information that differs from zero when group of attributes are complementary, which allows to prioritize causes with respect to irrelevant and effect variables.

However trivariate settings may present strong problems of indeterminacy, too. Think for instance to a fully connected triplet made of two causes and one common effect. In this case the lack of independency makes ineffective the adoption of conditional independency tests or interaction measures to infer the direction of the arrows. As stressed in (Guyon et al., 2007) when there are no independencies, the direction of the arrows can be anything. Though a possible remedy to indeterminacy comes from the use of additional instrumental variables (IV) (Bowden and Turkington, 1984), this strategy is not always feasible in real settings with lack of a priori knowledge.

This paper focuses on the definition of a data-dependent measure able to reduce the statistical indistinguishability of completely and linearly connected triplets. In particular, we propose a modification of the covariance formula of a structural equation model (Bollen, 1989; Mulaik, 2009) which results in a statistic taking opposite signs for different causal patterns when the unexplained variations of the variables are of the same magnitude. Though this assumption

Bontempi G. (2013).
A Statistic Criterion for Reducing Indeterminacy in Linear Causal Modeling.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 159-166
DOI: 10.5220/0004254301590166

could appear as too limiting, our rationale is that assumptions of comparable strength (e.g. the existence of unshielded colliders) have been commonly used so far in causal inference. We expect that this alternative approach could shed additional light on the issue of causality in the perspective of extending it to more general configurations.

For this reason the paper proposes also a Relevance Causal (RC) inference algorithm which integrates the proposed causal measure with a relevance measure to prioritize direct causes for a given target variable. In order to assess the effectiveness of the algorithm with respect to state-of-the-art algorithms a set of experiments aiming to infer linear and nonlinear networks from observed data are carried out. The experimental comparison with state-of-the-art techniques shows that such approach is promising for reducing indeterminacy in causal inference.

2 COVARIANCE OF A LINEARLY CONNECTED TRIPLET

The use of directed acyclic graphs (DAG) to encode causal dependencies and independencies is common to the two most known formalisms for causal modeling (Anderson and Vastage, 2004): Bayesian networks and structural equation models (SEM). These formalisms can accommodate both nonlinear and linear causal relationships. Here we will restrict our attention to the linear causal structure represented in Figure 1 where the variables $\mathbf{x}_1$ and $\mathbf{x}_2$ are causes of the random variable $\mathbf{x}_3$. Since a DAG can always be translated into a set of recursive structural equations, this linear dependency can be written as

$$\begin{cases} \mathbf{x}_1 &= \mathbf{w}_1 \\ \mathbf{x}_2 &= b_1\mathbf{x}_1 + \mathbf{w}_2 \\ \mathbf{x}_3 &= b_3\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{w}_3 \end{cases} \quad (1)$$

where it is assumed that each variable has mean 0, the $b_i \neq 0$ are also known as structural coefficients and the disturbances, supposed to be independent, are designated by $\mathbf{w}$. This set of equations can be put in the matrix form

$$\mathbf{x} = A\mathbf{x} + \mathbf{w} \quad (2)$$

where $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]^T$,

$$A = \begin{pmatrix} 0 & 0 & 0 \\ b_1 & 0 & 0 \\ b_3 & b_2 & 0 \end{pmatrix}$$

and $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3]^T$. The multivariate variance-covariance matrix (Mulaik, 2009) has no zero entries

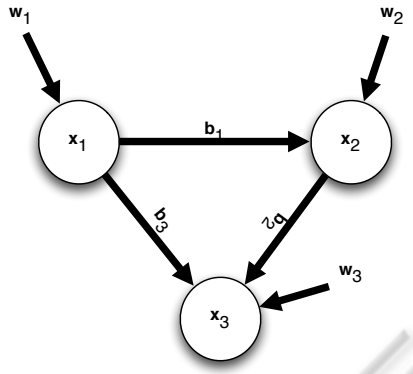

Figure 1: Collider pattern: completely connected triplet where the variable $\mathbf{x}_3$ is a common effect of $\mathbf{x}_1$ and $\mathbf{x}_2$.

and is given by

$$\Sigma = (I-A)^{-1}G((I-A)^T)^{-1} = \quad (3)$$

$$= \begin{bmatrix} \sigma_1^2 & b_1\sigma_1^2 & b_3\sigma_1^2 + b_1b_2\sigma_1^2 \\ b_1\sigma_1^2 & b_1^2\sigma_1^2 + \sigma_2^2 & b_1b_3\sigma_1^2 + b_2(b_1^2\sigma_1^2 + \sigma_2^2) \\ b_3\sigma_1^2 + b_1b_2\sigma_1^2 & b_1b_3\sigma_1^2 + b_2(b_1^2\sigma_1^2 + \sigma_2^2) & (b_1^2\sigma_1^2 + \sigma_2^2)b_2^2 + 2b_1b_2b_3\sigma_1^2 + b_3^2\sigma_1^2 + \sigma_3^2 \end{bmatrix} \quad (4)$$

where I is the identity matrix and

$$G = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

is the diagonal covariance matrix of the disturbances.

It is worthy noting here that the lack of zero entries in the covariance matrix (as well as in the inverse) illustrate the lack of conditional or unconditional independencies in the data. Constraint-based approaches (Spirtes et al., 2000) which rely on independence tests to retrieve the v-structure are consequently useless in this context. In the following section we will discuss whether SEM techniques can tackle such case.

3 INDETERMINACY IN A CONNECTED TRIPLET

Structural equation modeling techniques for causal inference proceed by 1) making some assumptions on the structure underlying the data, 2) perform the related parameter estimation, usually based on maximum likelihood and 3) assessing by significance testing the discrepancy between the sample covariance matrix and the covariance matrix implied by the hypothesis.

This section shows that, as known in litterature (Stelzl, 1986; Hershberger, 2006; Mulaik, 2009), conventional SEM is not able to reconstruct the right directionality of the connections in a completely connected triplet. Let us observe a set of data generated according to the dependency illustrated in Figure 1 and described algebrically by the set of structural equations (1). Suppose we want to test two alternative hypothesis, represented by the two directed graphs in Figure 2a and Figure 2b, respectively. Note that the hypothesis of Figure 2a is correct while the hypothesis illustrated by Figure 2b inverses the directionality of the link between $\mathbf{x}_2$ and $\mathbf{x}_3$ and consequently misses the causal role of the variable $\mathbf{x}_2$. Let us consider the following question: is it possible to discriminate between the structures 1 and 2 by simply relying on parameter estimation (in this case regression fitting) according to the hypothesized dependencies? The answer is unfortunately negative. Suppose we assess the hypothesis 1, by performing the two linear fittings implied by the hypothesis itself

$$\begin{cases} \mathbf{x}_2 = \hat{b}_1 \mathbf{x}_1 + \mathbf{w}_2 \\ \mathbf{x}_3 = \hat{b}_3 \mathbf{x}_1 + \hat{b}_2 \mathbf{x}_2 + \mathbf{w}_3 \end{cases}$$

where (Graybill, 1976)

$$\hat{b}_1 = \Sigma_{12}/\Sigma_{11} = b_1,$$

$$\begin{bmatrix} \hat{b}_3 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{13} \\ \Sigma_{23} \end{bmatrix} = \begin{bmatrix} b_3 \\ b_2 \end{bmatrix}$$

Since the above estimators are unbiased, if we compute the triplet covariance matrix by plugging the above estimates within the formula (3) we obtain again the variance in (4).

Let us consider now the second hypothesis (Figure 2b) and perform the two least-squares fittings

$$\begin{cases} \mathbf{x}_2 = \hat{b}_1 \mathbf{x}_1 + \hat{b}_2 \mathbf{x}_3 + \mathbf{w}_2 \\ \mathbf{x}_3 = \hat{b}_3 \mathbf{x}_1 + \mathbf{w}_3 \end{cases}$$

where the estimates are returned by

$$\hat{b}_3 = \Sigma_{13}/\Sigma_{11} = b_3 + b_1 b_2,$$

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{13} & \Sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{12} \\ \Sigma_{23} \end{bmatrix} = \begin{bmatrix} \frac{(b_1 * \sigma_3^2 - b_2 b_3 \sigma_2^2)}{b_2^2 \sigma_2^2 + \sigma_3^2} \\ \frac{b_2 \sigma_2^2}{b_2^2 \sigma_2^2 + \sigma_3^2)} \end{bmatrix}$$

Standard results give also the variance of the residuals. For instance the variance of $\mathbf{w}_2$ is returned by

$$\hat{\sigma}_2 = \Sigma_{22} - \begin{bmatrix} \Sigma_{12} \\ \Sigma_{23} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{13} \\ \Sigma_{13} & \Sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{12} \\ \Sigma_{23} \end{bmatrix}$$

We remark that, though the estimation of the parameters differs from the real structural coefficients, if we compute the complete covariance matrix by using (3) where

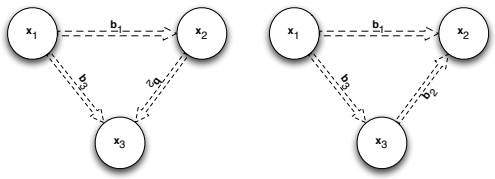

Figure 2: a) Hypothesis 1. b) Hypothesis 2.

$$\hat{A} = \begin{bmatrix} 0 & 0 & 0 \\ \hat{b}_1 & 0 & \hat{b}_2 \\ \hat{b}_3 & 0 & 0 \end{bmatrix}, \quad \hat{G} = \begin{bmatrix} \hat{\sigma_1}^2 & 0 & 0 \\ 0 & \hat{\sigma_2}^2 & 0 \\ 0 & 0 & \hat{\sigma}_3^2 \end{bmatrix}$$

we obtain again the expression (4). In other terms fitting different causal structures to the connected triplet does not allow to distinguish between the configuration in Figure 2a and the one in Figure 2b.

4 A CRITERION TO DETECT CAUSAL ASYMMETRY

A main characteristic of a causal relationship is its asymmetry. For this reason, if we wish to infer causal directionality from observational data we need to define discriminant criteria able to distinguish causes from effects. Let us suppose we want to discriminate between the causal patterns in Figure 1 where both $\mathbf{x}_1$ and $\mathbf{x}_2$ are direct causes of $\mathbf{x}_3$ and alternative patterns like the ones in Figure 3a and Figure 3b. As we have seen in the previous section, the conventional SEM procedure does not allow to distinguish between different patterns. What we propose here is an alternative criterion able to perform such distinction.

The computation of our criterion requires the fitting of the two hypothetical structures in Figures 2a and 2b to the data, as done in the previous section. What is different is that, instead of computing the term (3), we consider the term

$$S = (I - A)^{-1}((I - A)^T)^{-1}. \tag{5}$$

Let us see the impact of such modification in the detection of causality by analyzing in the following sections three different causal patterns. In all cases we will make the assumption that $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$, i.e. that the unexplained variations of the variables are of comparable magnitude. Though we are aware that this assumption is quite specific, some considerations are worthy to be made. So far, most of the approaches of causal inference from data have relied on similar, if not stronger, assumptions like postulating the existence of unshielded colliders. At the same time the

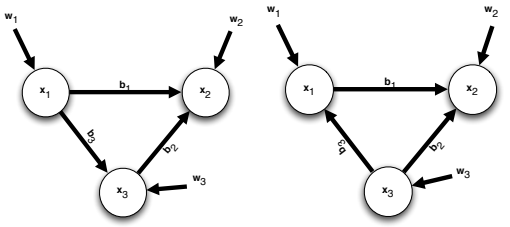

Figure 3: a) Chain pattern, completely connected triplet where the variable $\mathbf{x}_2$ is the common effect of $\mathbf{x}_1$ and $\mathbf{x}_3$. b) Fork pattern: completely connected triplet where the variables $\mathbf{x}_2$ and $\mathbf{x}_1$ have the common cause $\mathbf{x}_3$.

following derivation is expected to shed a new light on the issue of causality with the aim of applying it to more general configurations.

4.1 Collider Pattern

Let us suppose that data are generated according to the structure in Figure 1 where the node $\mathbf{x}_3$ is a collider.

If we fit the hypothesis 1 to the data and we compute the term (5) we obtain

$$\hat{S}_1 = (I-A_1)^{-1}((I-A_1)^T)^{-1} = \begin{bmatrix} 1 & b_1 & (b_3+b_1b_2) \\ b_1 & b_1^2+1 & b_1(b_3+b_1b2)+b_2] \\ (b_3+b_1b_2) & b_1(b_3+b_1b2)+b_2] & (b_3+b_1b_2)^2+1+b_2^2 \end{bmatrix}$$

If we fit the hypothesis 2 to data and we compute the term (5) we obtain

$$\hat{S}_2 = (I-A_2)^{-1}((I-A_2)^T)^{-1} = \begin{bmatrix} 1 & b_1 & b_3+b_1b_2 \\ b_1 & \frac{b_2^2}{(b_2^2+1)^2}+b_1^2+1 & \frac{b_2}{(b_2^2+1)}+b_1(b_3+b_1b_2) \\ b_3+b_1b_2 & \frac{b_2}{(b_2^2+1)}+b_1(b_3+b_1b_2) & (b_3+b_1b_2)^2+1 \end{bmatrix}$$

Let us denote by $S[i,j]$ the ijth element of a matrix S. Since $\forall i\ b_i \neq 0$, it follows that the quantity

$$C(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = = \hat{S}_1[3,3]-\hat{S}_2[3,3]+\left(\hat{S}_1[2,2]-\hat{S}_2[2,2]\right) = = \frac{{b_2}^4\left({b_2}^2+2\right)}{\left({b_2}^2+1\right)^2} \quad (6)$$

is greater than zero for any sign of the structural coefficients. Interestingly enough, the sign is preserved for $\sigma_1=\sigma_2=\sigma_3$ also when the direction of the link between $\mathbf{x}_1$ and $\mathbf{x}_2$ is inverted (see (18).

4.2 Chain Pattern

We suppose here that the data have been generated by the triplet in Figure 3, where $\mathbf{x}_3$ is part of the chain pattern $\mathbf{x}_1 \rightarrow \mathbf{x}_3 \rightarrow \mathbf{x}_2$. This configuration is represented by the matrix

$$A = \begin{pmatrix} 0 & 0 & 0 \\ b_1 & 0 & b_2 \\ b_3 & 0 & 0 \end{pmatrix}$$

Let us proceed by computing the quantity C in (6) for such generative model under the assumption $\sigma_1 = \sigma_2 = \sigma_3$. For the sake of space we will report here only the components of the submatrices $\hat{S}_1[2:3,2:3]$ and $\hat{S}_2[2:3,2:3]$. If data have been generated according to the structure in Figure 3 and we fit the hypothesis 1 we obtain

$$\hat{S}_1[2:3,2:3] = \begin{bmatrix} (b_1+b_2b_3)^2+1 & \frac{b_2}{(b_2^2+1)}+b_3(b_1+b_2b_3) \\ \frac{b_2}{(b_2^2+1)}+b_3(b_1+b_2b_3) & \frac{b_2^2}{(b_2^2+1)^2}+b_3^2+1 \end{bmatrix} \quad (7)$$

If data have been generated according to the structure in Figure 3 and we fit the hypothesis 2 we obtain

$$\hat{S}_2[2:3,2:3] = \begin{bmatrix} (b_1+b_2b_3)^2+b_2^2+1 & b_2+b_3(b_1+b_2b_3) \\ b_2+b_3(b_1+b_2b_3) & b_3^2+1 \end{bmatrix} \quad (8)$$

It follows that

$$C(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = = \hat{S}_1[3,3]-\hat{S}_2[3,3]+\left(\hat{S}_1[2,2]-\hat{S}_2[2,2]\right) = = -\frac{(b_2^4(b_2^2+2))}{(b_2^2+1)^2} \quad (9)$$

This term is less than zero whatever the sign of the structural coefficients b_i in Figure 3.

Note that we do not discuss here the configuration with the edge pointing from $\mathbf{x}_2$ to $\mathbf{x}_1$ since this is a cyclic one.

4.3 Fork Pattern

Suppose now that observations are generated by the triplet in Figure 3b, corresponding to the matrix

$$A = \begin{pmatrix} 0 & 0 & b_3 \\ b_1 & 0 & b_2 \\ 0 & 0 & 0 \end{pmatrix}$$

Like in the previous section we report the components of the submatrices $\hat{S}_1[2:3,2:3]$ and $\hat{S}_2[2:3,2:3]$:

$$\hat{S}_1[2:3,2:3] = \\ = \begin{bmatrix} \frac{(b_1 b_3^2 + b_2 b_3 + b_1)^2}{(b_3^2+1)^2} + 1 & \\ \frac{b_2}{(b_2^2+b_3^2+1)} + \frac{(b_3(b_1 b_3^2 + b_2 b_3 + b_1))}{(b_3^2+1)^2} & \\ \frac{b_2}{(b_2^2+b_3^2+1)} + \frac{(b_3(b_1 b_3^2 + b_2 b_3 + b_1))}{(b_3^2+1)^2} & \\ \frac{b_2^2}{(b_2^2+b_3^2+1)^2} + \frac{b_3^2}{(b_3^2+1)^2} + 1 & \end{bmatrix} \quad (10)$$

$$\hat{S}_2[2:3,2:3] = \begin{bmatrix} (b_1 + b_2 b_3)^2 + b_2^2 + 1 & \\ b_2 + b_3(b_1 + b_2 b_3) & \\ b_2 + b_3(b_1 + b_2 b_3) & \\ b_3^2 + 1 & \end{bmatrix} \quad (11)$$

It follows that

$$C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \\ = \hat{S}_1[3,3] - \hat{S}_2[3,3] + \left(\hat{S}_1[2,2] - \hat{S}_2[2,2]\right) = \\ = b_2^2 \left(\frac{1}{(b_2^2 + b_3^2 + 1)^2} - 1 \right) \quad (12)$$

This term is less than zero whatever the sign of the structural coefficients $b_i \neq 0$ in Figure 3b. In the supplementary material we compute the value of C when the direction of the link between $\mathbf{x}_1$ and $\mathbf{x}_2$ is reversed (matrix (19) in the supplement). From (22) we obtain that this value remains negative when $(b_2^2 + b_3^2) > b_1^2$, for instance when the absolute value of one of the coefficients associated to the edges leaving $\mathbf{x}_3$ is bigger than $|b_1|$. In plain words if the cause-effect relationship between $\mathbf{x}_3$ and the other variables is strong enough, the statistics C takes a negative value.

The equations (6), (9) and (12) show that the computation of the quantity C on the basis of observational data only can help in discriminating between the collider configuration in Figure 1 where the nodes $\mathbf{x}_1$ and $\mathbf{x}_2$ are direct causes of $\mathbf{x}_3$ ($C > 0$) and non collider configurations (i.e. fork or chain) ($C < 0$) in Figure 3a and 3b.

In other terms, given a completely connected triplet of variables, the quantity $C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ returns useful information about the causal role of $\mathbf{x}_1$ and $\mathbf{x}_2$ with respect to $\mathbf{x}_3$ whatever is the strength or the direction of the link between $\mathbf{x}_1$ and $\mathbf{x}_2$.

5 A RELEVANCE CAUSAL ALGORITHM TO INFER DIRECTIONALITY

The properties of the quantity C encourage its use in an algorithm to infer directionality from observational data. We propose then a RC (Relevance Causal) algorithm for linear causal modeling inspired to the mIMR causal filter selection algorithm (Bontempi and Meyer, 2010). The mIMR algorithm is characterized by two terms, a relevance term, assessing the relevance of each input variable with respect to a target variable and a causation term, aiming to prioritize causal variables by minimizing the interaction of triplets of variables. The causation term is designed in order to reward variables which belong to a collider pattern and penalize variables within a fork pattern. Let us suppose that we want to identify the set of causes of a target variable $\mathbf{y}$ among a set $\mathbf{X}$ of inputs. The mIMR is a forward selection algorithm which given a set $\mathbf{X}_S$ of d already selected variables, updates this set by adding the $d+1$th variable which satisfies

$$\mathbf{x}^*_{d+1} = \arg \max_{\mathbf{x}_k \in \mathbf{X} - \mathbf{X}_S} \Big[(1-\lambda) I(\mathbf{x}_k; \mathbf{y}) \\ - \frac{\lambda}{d} \sum_{\mathbf{x}_i \in \mathbf{X}_S} I(\mathbf{x}_i; \mathbf{x}_k; \mathbf{y}) \Big] \quad (13)$$

where $I(\mathbf{x}_k; \mathbf{y})$ denotes the mutual information between $\mathbf{x}_k$ and $\mathbf{y}$, $I(\mathbf{x}_i; \mathbf{x}_k; \mathbf{y})$ denotes the interaction information and the coefficient $\lambda \in [0,1]$ is used to weight the mutual information and the interaction term.

As discussed previously, this algorithm might suffer of bad performance when common causes are directly connected since the interaction term $I(\mathbf{x}_i; \mathbf{x}_k; \mathbf{y})$ could take positive values for a v-structure $\mathbf{x}_i \rightarrow \mathbf{y} \leftarrow \mathbf{x}_k$. For that reason we propose to replace the interaction term (to be minimized) with the criterion C (to be maximized) to infer causal dependency from observed data also in presence of completed connected triplets. The resulting algorithm is a reformulation of the mIMR where the update formula is now

$$\mathbf{x}^*_{d+1} = \arg \max_{\mathbf{x}_k \in \mathbf{X} - \mathbf{X}_S} \Big[(1-\lambda) R(\{\mathbf{X}_S, \mathbf{x}_k\}; \mathbf{y}) + \\ \frac{\lambda}{d} \sum_{\mathbf{x}_i \in \mathbf{X}_S} C(\mathbf{x}_i; \mathbf{x}_k; \mathbf{y}) \Big] \quad (14)$$

where $\lambda \in [0,1]$ weights the R and the C contribution, the R term quantifies the relevance of the subset $\{\mathbf{X}_S, \mathbf{x}_k\}$ and the C term quantifies the causal role of an input $\mathbf{x}_k$ with respect to the set of selected variables $\mathbf{x}_i \in \mathbf{X}_S$.

The proposed RC algorithm is then a forward selection algorithm which sequentially adds variables according to the update rule (14). Note that for $\lambda = 0$ the algorithm boils down to a conventional forward selection wrapper which assesses the subsets according to the measure R. The RC algorithm is initialized by selecting the couple of variables $\{\mathbf{x}_i, \mathbf{x}_j\}$ maximizing the quantity

$$(1-\lambda)R(\{\mathbf{x}_i, \mathbf{x}_j\}; \mathbf{y}) + \frac{\lambda}{d} \sum_{\mathbf{x}_i \in \mathbf{X}_S} C(\mathbf{x}_i; \mathbf{x}_j; \mathbf{y})$$

In the implementation used in the experimental section, we adopt a linear leave-one-out measure to quantify the relevance of a subset, i.e. $R(\mathbf{X}, \mathbf{y})$ is set equal to the negative of linear leave-one-out mean-squared-error of the regression with input $\mathbf{X}$ and target $\mathbf{y}$. Also in order to have comparable values for the R and the C terms, at each step these quantities are normalized over the interval $[0,1]$ before performing their weighted sum.

6 EXPERIMENTS

In this section we assess the efficacy of the RC algorithm by performing a set of causal network inference experiments. The aim of the experiment is to reverse engineer both linear and nonlinear scale-free causal networks, i.e. networks where the distribution of the degree follows a power law, from a limited amount of observational data. We consider a set of networks with a large number $n = 5000$ of nodes and where the degree α of the power law ranges between 2.1 and 3. The inference is done on the basis of a small amount of $N = 200$ observations. The structural coefficients of the linear dependencies have an absolute value distributed uniformly between 0.5 and 0.8, and the measurement error follows a standard Normal distribution. Nonlinear networks are obtained by transforming the linear dependencies between nodes with a sigmoid function.

We compare the accuracy of several algorithms in terms of the mean F-measure (the higher, the better) averaged over 10 runs and over all the nodes with a number of parents and children larger equal than two. The F-measure, also known as balanced F-score, is the weighted harmonic mean of precision and recall and is conventionally used to provide a compact measure of the quality of a network inference algorithm. We considered the following algorithms for comparison: the IAMB algorithm (Tsamardinos et al., 2003) implemented by the Causal Explorer software (Aliferis et al., 2003) which estimates for a given variable the set of variables belonging to its Markov blanket, the mIMR (Bontempi and Meyer, 2010) algorithm, the mRMR (Peng et al., 2005) algorithm and three versions of the RC algorithm with three different values $\lambda = 0, 0.5, 1$. Note that the *RC* algorithm with $\lambda = 0$ boils down to a conventional wrapper algorithm based on the leave-one-out assessment of the variables' subsets.

We also remark that the RC algorithms aims to return for a given node a prioritization of the other nodes according to their causal role while the Causal Explorer implementation of IAMB returns a specific subset (for a given pvalue). For the sake of comparison, we decided to compute the F-measure by setting the number of putative causes to the number of variables returned by IAMB.

Tables 1 and 2 report the average F-measures for different values of α in the linear and nonlinear case, respectively.

The results show the potential of the criterion C and of the RC algorithm in network inference tasks where dependencies between parents are frequent because of direct links or common ancestors. According to the F-measures reported in the Tables the RC accuracy with $\lambda = 0.5$ and $\lambda = 1$ is coherently better than the ones of mIMR, mRMR and IAMB algorithms for all the considered degrees distributions. However the most striking result is the clear improvement with respect to a conventional wrapper approach which targets only prediction accuracy ($\lambda = 0$) when a causal criterion C is taken into account together with a predictive one ($\lambda = 0.5$). These results confirm previous results (Bontempi and Meyer, 2010; Bontempi et al., 2011) putting into evidence that an effective causal inference task should combine a relevance criterion targeting prediction accuracy with a causal term able to prioritize direct cause and penalize effects.

7 CONCLUSIONS

Causal inference from complex large dimensional data is taking a growing importance in machine learning and knowledge discovery. Currently, most of the existing algorithms are limited by the fact that the discovery of causal directionality is submitted to the detection of a limited set of distinguishable patterns, like unshielded colliders. However the scarcity of data and the intricacy of dependencies in networks could make the detection of such patterns so rare that the resulting

Table 1: Linear case: F-measure (averaged over all nodes with a number of parents and children ≥ 2 and over 10 runs) of the accuracy of the inferred networks on the basis of $N = 100$ observations.

α	IAMB	mIMR	mRMR	RC_0	$RC_{0.5}$	RC_1
2.2	0.375	0.324	0.319	0.386	0.421	0.375
2.3	0.378	0.337	0.333	0.387	0.437	0.401
2.4	0.376	0.342	0.342	0.385	0.441	0.414
2.5	0.348	0.322	0.313	0.358	0.422	0.413
2.6	0.347	0.318	0.311	0.355	0.432	0.414
2.7	0.344	0.321	0.311	0.352	0.424	0.423
2.8	0.324	0.304	0.293	0.334	0.424	0.422
2.9	0.342	0.333	0.321	0.353	0.448	0.459
3.0	0.321	0.319	0.297	0.326	0.426	0.448

Table 2: Nonlinear case: F-measure (averaged over all nodes with a number of parents and children ≥ 2 and over 10 runs) of the accuracy of the inferred network on the basis of $N = 100$ observations.

α	IAMB	mIMR	mRMR	RC_0	$RC_{0.5}$	RC_1
2.2	0.312	0.310	0.304	0.314	0.356	0.324
2.3	0.317	0.328	0.316	0.320	0.375	0.349
2.4	0.304	0.317	0.304	0.306	0.366	0.351
2.5	0.321	0.327	0.328	0.325	0.379	0.359
2.6	0.306	0.325	0.306	0.309	0.379	0.365
2.7	0.313	0.319	0.303	0.316	0.380	0.359
2.8	0.297	0.326	0.300	0.300	0.392	0.382
2.9	0.310	0.329	0.313	0.313	0.389	0.377
3.0	0.299	0.324	0.300	0.303	0.399	0.392

precision would be unacceptable. This paper shows that it is possible to identify new statistical measures helping in reducing indistinguishability under the assumption of equal variances of the unexplained variations of the three variables. Though this assumption could be questioned, we deem that it is important to define new statistics to help discriminating between causal structures for completely connected triplets in linear causal modeling. Future work will focus on assessing whether such statistic is useful in reducing indeterminacy also when the assumption of equal variance is not satisfied.

REFERENCES

Aliferis, C., Tsamardinos, I., and Statnikov, A. (2003). Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *The 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03)*.

Anderson, R. and Vastage, G. (2004). Causal modeling alternatives in operations research: overview and application. *Eurpean Journal of Operational Research*, 156:92–109.

Bollen, K. (1989). *Structural equations with latent variables*. John Wiley and Sons.

Bontempi, G., Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Quackenbush, J. (2011). Multiple-input multiple-output causal strategies for gene selection. *BMC bioinformatics*, 12(1):458.

Bontempi, G. and Meyer, P. (2010). Causal filter selection in microarray data. In *Proceeding of the ICML2010 conference*.

Bowden, R. and Turkington, D. (1984). *Instrumental Variables*. Cambridge University Press.

Brown, G. (2009). A new perspective for information theoretic feature selection. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Graybill, F. (1976). *Theory and Application of the Linear Model*. Duxbury Press.

Guyon, I., Aliferis, C., and Elisseeff, A. (2007). *Computational Methods of Feature Selection*, chapter Causal Feature Selection, pages 63–86. Chapman and Hall.

Hershberger, S. (2006). *Structural equation modeling: a second course*, chapter The problems of equivalent structural models, pages 13–41. Springer.

Janzing, D., Hoyer, P. O., and Scholkopf, B. (2010). Telling cause from effect based on high-dimensional observations. In *Proceeding of the ICML2010 conference*.

Janzing, D., Sgouritsa, E., Stegle, O., Peters, J., and Scholkopf, B. (2011). Detecting low-complexity unobserved causes. In *Conference on Uncertainty in Artificial Intelligence (UAI2011)*.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models*. The MIT Press.

Mulaik, S. (2009). *Linear Causal Modelling with Structural Equations*. CRC Press.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency,max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. Springer Verlag, Berlin.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, 21:309?331.

Tsamardinos, I., Aliferis, C., and Statnikov, A. (2003). Algorithms for large scale markov blanket discovery. In *Proceedings of the 16th International FLAIRS Conference (FLAIRS 2003)*.

Watkinson, J., Liang, K., Wang, X., Zheng, T., and Anastassiou, D. (2009). Inference of regulatory gene interactions from expression data using three-way mutual information. *Annals of N.Y. Academy of Sciences*, 1158:302–313.

APPENDIX

Let

$$A = \begin{bmatrix} 0 & b_1 & 0 \\ 0 & 0 & 0 \\ b_3 & b_2 & 0 \end{bmatrix} \tag{15}$$

be the matrix associated to the collider pattern with a edge heading from $\mathbf{x}_2$ to $\mathbf{x}_1$. We compute the quantity C for such generative model under the assumption $\sigma_1 = \sigma_2 = \sigma_3$.

If data have been generated according to the structure (15) and we fit the hypothesis 1 we obtain

$$\hat{S}_1[2:3,2:3] = \begin{bmatrix} \frac{b_1^2}{(b_1^2+1)^2}+1 & b_2+\frac{(b_1(b_3b_1^2+b_2b_1+b_3))}{(b_1^2+1)^2} \\ b_2+\frac{(b_1(b_3b_1^2+b_2b_1+b_3))}{(b_1^2+1)^2} & \frac{(b_3b_1^2+b_2b_1+b_3)^2}{(b_1^2+1)^2}+b_2^2+1 \end{bmatrix} \tag{16}$$

If we fit the hypothesis 2 we obtain

$$\hat{S}_2[2:3,2:3] = \begin{bmatrix} \frac{b_2^2}{(b_1^2+b_2^2+1)^2}+\frac{b_1^2}{(b_1^2+1)^2}+1 & \frac{b_2}{(b_1^2+b_2^2+1)}+\frac{(b_1(b_3b_1^2+b_2b_1+b_3))}{(b_1^2+1)^2} \\ \frac{b_2}{(b_1^2+b_2^2+1)}+\frac{(b_1(b_3b_1^2+b_2b_1+b_3))}{(b_1^2+1)^2} & \frac{(b_3b_1^2+b_2b_1+b_3)^2}{(b_1^2+1)^2}+1 \end{bmatrix} \tag{17}$$

It follows that

$$\begin{aligned} C(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \\ &= \hat{S}_1[3,3]-\hat{S}_2[3,3]+\left(\hat{S}_1[2,2]-\hat{S}_2[2,2]\right) = \\ &= b_2^2-\frac{b_2^2}{(b_1^2+b_2^2+1)^2} > 0 \end{aligned} \tag{18}$$

In other words the sign is positive also in case of a link from $\mathbf{x}_2$ to $\mathbf{x}_1$.

Let us consider now the fork pattern described by the matrix

$$A = \begin{bmatrix} 0 & b_1 & b_3 \\ 0 & 0 & b_2 \\ 0 & 0 & 0 \end{bmatrix} \tag{19}$$

If data have been generated according to the structure (19) and we fit the hypothesis 1 we obtain

$$\hat{S}_1[2:3,2:3] = \begin{bmatrix} \frac{(b_1b_2^2+b_3b_2+b_1)^2}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2}+1 & \frac{(b_2-b_1b_3)}{(b_2^2+b_3^2+1)}+\frac{((b_3+b_1b_2)(b_1b_2^2+b_3b_2+b_1))}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2} \\ \frac{(b_2-b_1b_3)}{(b_2^2+b_3^2+1)}+\frac{((b_3+b_1b_2)(b_1b_2^2+b_3b_2+b_1))}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2} & \frac{(b_3+b_1b_2)^2}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2}+\frac{(b_2-b_1b_3)^2}{(b_2^2+b_3^2+1)^2}+1 \end{bmatrix} \tag{20}$$

If we fit the hypothesis 2 we obtain

$$\hat{S}_2[2:3,2:3] = \begin{bmatrix} \frac{(b_2-b_1b_3)^2}{(b_1^2+1)^2}+\frac{(b_1b_2^2+b_3b_2+b_1)^2}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2}+1 & \frac{(b_2-b_1b_3)}{(b_1^2+1)}+\frac{((b_3+b_1b_2)(b_1b_2^2+b_3b_2+b_1))}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2} \\ \frac{(b_2-b_1b_3)}{(b_1^2+1)}+\frac{((b_3+b_1b_2)(b_1b_2^2+b_3b_2+b_1))}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2} & \frac{(b_3+b_1b_2)^2}{(b_1^2b_2^2+b_1^2+2b_1b_2b_3+b_3^2+1)^2}+1 \end{bmatrix} \tag{21}$$

It follows that

$$\begin{aligned} C(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \\ &= \hat{S}_1[3,3]-\hat{S}_2[3,3]+\left(\hat{S}_1[2,2]-\hat{S}_2[2,2]\right) = \\ &= \frac{(b_2-b_1b_3)^2}{(b_2^2+b_3^2+1)^2}-\frac{(b_2-b_1b_3)^2}{(b_1^2+1)^2} = \\ &= (b_2-b_1b_3)^2\left(\frac{1}{(b_2^2+b_3^2+1)^2}-\frac{1}{(b_1^2+1)^2}\right) \end{aligned} \tag{22}$$

Note that this quantity is negative when $(b_2^2+b_3^2) > b_1^2$.