An Analytic and Empirical Comparison of Two Methods for Discovering Probabilistic Causal Relationships

Donato Malerba, Giovanni Semeraro and Floriana Esposito
Dipartimento di Informatica, Università degli Studi di Bari
via Orabona 4, 70126 Bari, Italy
{malerbad | semeraro | esposito}@vm.csata.it

Abstract. The discovery of causal relationships from empirical data is an important problem in machine learning. In this paper the attention is focused on the inference of probabilistic causal relationships, for which two different approaches, namely Glymour et al.'s approach based on constraints on correlations and Pearl and Verma's approach based on conditional independencies, have been proposed. These methods differ both in the kind of constraints they consider while selecting a causal model and in the way they search the model which better fits to the sample data. Preliminary experiments show that they are complementary in several aspects. Moreover, the method of conditional independence can be easily extended to the case in which variables have a nominal or ordinal domain. In this case, symbolic learning algorithms can be exploited in order to derive the causal law from the causal model.

1 Introduction

Both a fortune-teller and a financial planner can make a forecast on the result of an investment in stocks. The former will use a pack of cards while the latter will base his/her prediction on his/her knowledge about the present economical situation as well as the economical, social and psychological processes which control the state of the Stock Exchange. Although it may happen that the fortune-teller's prediction is more accurate than that made by a financial planner, few people would be inclined to entrust a fortune-teller with their own savings. The reason for such a distrust is that a fortune-teller relates the result of the investment to the obscure meaning of the arrangement of the cards while the financial expert is able to *explain* why he/she made a certain prediction.

Generally speaking, the *explanatory* knowledge of a phenomenon allows us not only to make a forecast just like the *declarative* knowledge, but also to explain the reasoning followed in order to reach some conclusions. Since causality plays an important role in human understanding, it can be fairly stated that causal knowledge is a fundamental part of any intelligent system. For instance, Steels [12] pointed out that the major problem of the first expert systems was their inability to provide effective explanations of why an observed phenomenon happened rather than very simple explanations based on the chain of heuristic rules used to reach a conclusion.

In recent years, the exigency of disposing of deep knowledge on the application domain has been raised in machine learning as well. In particular, the importance of disposing of a domain theory which controls the inferential process of a learning system has been stressed, since it would be better to find generalizations which can explain the learned concepts themselves other than correctly classify new observations. The increasing interest towards explanation-based learning is just due to its ability of

generalizing concepts by using an appropriate domain theory [6, 3]. Nevertheless, the real applicability of this learning paradigm is strongly conditioned by the possibility of defining a domain theory, often comprising a *causal model*, which is complete, consistent and tractable.

As to the completeness property, the definition of a complete theory can be made easier by means of systems which can inductively discover causal relations from data, that is a theory which can explain every example from the domain theory under study.

A cause-effect relationship can be deterministic or probabilistic. A deterministic relationship can be established when the description of all the variables involved in a phenomenon is sufficiently detailed, while a probabilistic relationship exists when some relevant variables cannot be measured or when some measurements are not sufficiently accurate. As Suppes [13] has already emphasized, most of the causal relationships that humans use in their reasoning are probabilistic.

In this paper, we compare two statistical approaches to discovering causal relationships, namely Glymour et al.'s method of constraints on correlations [4, 11] and Pearl and Verma's method of conditional independencies [7]. Preliminary experiments indicate that these methods are complementary since they show very different results for diverse causal models. Moreover, these methods can only discover a causal dependence between variables in the model and not the causal law. When the variables in the model are numerical (interval or ratio level measurements), coefficients of the linear models can be estimated by means of regression analysis. However, the method based on conditional independence can also be applied to nominal or ordinal variables, in which case symbolic inductive learning algorithms can help to find the causal rules of the phenomenon. Thus, inductive learning algorithms together with statistical causal inference systems can be exploited in order to discover causal relationships which can be subsequently used as a part of the domain theory of an analytic learning system.

2 Preliminaries

A *causal model* is the abstraction of a set of cause-effect relationships from a statistical model in which we ignore the equations and most of the statistical assumptions. A causal model of a set of variables V can be represented by a directed graph, whose nodes are distinct elements of V and whose edges denote direct causal relationships between pairs of variables (see Figure 1). Henceforth, we will consider only the case of acyclic causal models, which can be represented by directed acyclic graphs (*dags*). Moreover, we assume

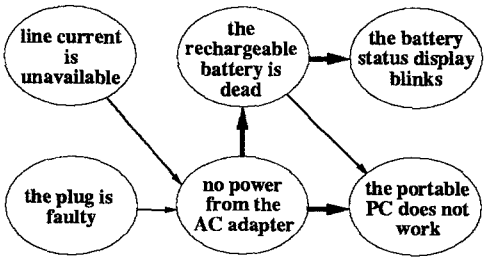

Figure 1. A causal model referring to power problems in a portable PC. Bold arrows indicate a trek between the two effects "the battery status display blinks" and "the portable PC does not work."

that the reader is familiar with some basic notions on dags, so we limit ourselves to introduce only the following concepts:

An *acyclic (open) path* is a path which contains no cycle. A *trek* between two distinct nodes v and w is a pair of acyclic paths, p and q, from a node u to v and w respectively, such that they intersect only in u. The trek is denoted by $<p,q>$ and u is called *source* of the trek. Bold arrows in Figure 1 show a trek between the nodes "the battery status display blinks" and "the portable PC does not work." The source of the trek is the variable "no power from the AC adapter." Henceforth, T_{uv} will denote the set of treks between u and v, while P_{uv} will denote the set of acyclic paths from u to v. Note that when u and v coincide, then $T_{uv}=P_{uv}$. For any node v of a dag, *indegree(v)* is equal to the number of edges directed into v, while *outdegree(v)* is equal to the number of edges directed out of v.

Definition 1 ***(stochastic causal theory)***

A *stochastic causal theory* (*SCT*) is a triple $(<V,E>, (\Omega,P), X)$ where:

1) $<V,E>$ is a causal model;
2) (Ω,P) is a probability space for the random variables in V. Every variable in V has a non-zero variance. If there is no trek between two variables $u,v\in V$, then u and v are statistically independent. Moreover, each variable $v\in V$ having a causal predecessor ($indegree(v)>0$) is associated with one *error variable* e_v which takes into account external sources of variance.
3) X is a set of independent equations in V. For each $v\in V$ such that $indegree(v)>0$ the equation:

$$v = f_v(w_{v1}, w_{v2}, ..., w_{vn}, e_v)$$

is a member of X, *where* $wv1, wv2, ..., w_{vn}$ are all variables adjacent to v.

A variable v is said to be *independent* iff $indegree(v)=0$, otherwise it is *dependent*. Dependent variables are the effects of some causes, and they are associated with some disturbance terms, that is, the error variables. These latter represent the increment by which any individual v may fall off the regression line:

$$v = f_v(w_{v1}, w_{v2}, ..., w_{vn}, 0)$$

According to definition 1, stochastic causal theories may differ in the form of the relationship between a set of causes (variables adjacent to v) and an effect (v itself), that is, in the form of f_v. SCTs in which f_v is a linear function are of peculiar interest.

Definition 2 ***(stochastic linear causal theory)***

A *stochastic linear causal theory* (*SLCT*) is a 4-tuple $(<V,E>, (\Omega,P), X, L)$ where:

1) $(<V,E>, (\Omega,P), X)$ is a stochastic causal theory;
2) L is a *labelling function* for the edges in E taking values in the set of non-null real numbers. For any edge e from u to v, we will denote $L(e)=a_{vu}$. The label of a path p, $L(p)$, is defined as the product of the labels of each edge in p. The label of a trek $t=<p,q>$ is defined as $L(t)=L(p)L(q)$.
3) X is a set of independent homogeneous linear equations in V. For each $v\in V$ such that $indegree(v)>0$ the equation:

$$v = \sum_{w\in Adj(v)} a_{vw}w + e_v$$

is a member of X, where $Adj(v)=\{w\in V \mid w \text{ is adjacent to } v\}$.

The label associated to an edge from u to v summarizes the total impact on v of a unit change in u after it has rippled through the causal model. Such an impact on v can be positive or negative according to the sign of the label. Consequently, given a path p from u to v, the associated label represents the total impact on v of a unit change of u when only the causal relationships in p are considered. Figure 2 shows an example of labelled causal model and the set of linear equations derived from it. Here λ is a *latent* variable, that is a not measured variable which represents a cause of the phenomenon described by the causal model and has a possible theoretical explanation. Also error variables e_i are not measured, but differently from latent variables they do not carry with them any theoretical interpretation and do not represent a common cause of two or more variables. All the other variables, denoted by Latin characters, are *measured* or *observed*, that is a sample is available for all them.

3 The Method of Constraints on Correlations

Explaining observed data is the main reason for which a theory is built. In the case of statistical linear theories, the aim is usually the explanation of correlations or covariances between random variables whose values have been collected in a sample. In particular, for stochastic linear causal theories, the goal is that of discovering causal relationships by exploiting the information coming from correlations between measured variables. However, it is generally raised the following question: how can a correlation (or a covariance), which is a symmetric piece of information, provide hints on the direction of causality? Of course, if we knew which variable temporally precedes which we could also define univocally the causal direction. When temporal information is unavailable, temporal constraints can be replaced by constraints on correlations. However, by using a simple correlation between two variables no one can discriminate the concomitant variation due to a common (hidden) cause from a real causal influence of a variable on the other one. On the contrary, the causal relationship between two variables can be determined by examining the covariances of a larger set of variables. For instance, if we are given two variables, x="line current is unavailable" and y="no power from the AC adapter", we can conclude that x causes y only if we consider a further *control* variable z, say "the plug is faulty", such that z is correlated to y but not to x (see Figure 1).

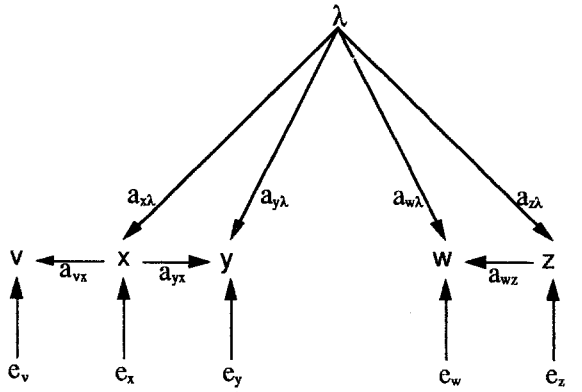

Figure 2. A labelled causal model and the set of linear equations derived from it.

Symmetrically, we can also conclude that z causes y by taking x as control variable.

Given a SLCT $T = (<V,E>, (\Omega,P), X, L)$, Glymour et al. [4, pp. 285-286] have proven the following formula for the covariance γ_{xy} between two variables x and y:

$$\gamma_{xy} = \sum_{w \in T_{xy}} L(t)\, \sigma^2_{s(t)}$$

where $\sigma^2_{s(t)}$ is the variance of the variable $s(t)$, the source of a trek t between x and y.

Note that, if variables are standardized apart, then $\sigma^2_v = 1$ for any variable $v \in V$ and the covariance γ_{xy} equals the correlation ρ_{xy}. Thus, the correlation between two measured variables is equal to the sum, over all treks connecting the variables, of the product of the coefficients corresponding to the directed edges in the trek [5]. Indeed, in the case of variables that are standardized apart, the label associated to an edge from u to v denotes the correlation between u and v.

Covariances computed according to the previous formula are said to be *implied* by the theory, in order to distinguish them from the sample covariances. Unfortunately, the computation of γ_{xy} for any two variables x and y requires knowledge about the edge labels and the variances of the independent variables. Instead of estimating coefficients and variances in order to compare the *values* of implied correlations with the corresponding values of sample correlations, Glymour et al. prefer to compare *constraints* on covariances (or correlations) implied by the model with constraints on covariances (or correlations) which hold in the data. Thus, the problem is to *search for the causal model which implies the set of constraints that best fits to the set of constraints which hold in the data.*

There are several kinds of constraints on covariances which could be taken into account, but some considerations on computational complexity drive Glymour et al. to consider only two kinds of constraints: *tetrad* and *partial equations*. They are enough to choose among several alternative causal models and can be easily tested as well.

Definition 3 ***(tetrad equation)***

Let $M=<V,E>$ be a causal model, and x, y, w, z four distinct measured variables in V. A *tetrad equation* between x, y, w, z is one of the following equations:

$$\rho_{xy} \cdot \rho_{wz} = \rho_{xw} \cdot \rho_{yz}$$
$$\rho_{xy} \cdot \rho_{wz} = \rho_{xz} \cdot \rho_{yw}$$
$$\rho_{xw} \cdot \rho_{yz} = \rho_{xz} \cdot \rho_{yw}$$

where ρ_{ij} is the correlation between i and j.

The above equations represent constraints on a foursome of variables, while *partial equations* define constraints on a triplet of variables.

Definition 4 ***(partial equation)***

Let $M=<V,E>$ be a causal model, and x, y, z three distinct measured variables in V. A *partial equation* between x, y, z is the following equation:

$$\rho_{xz} = \rho_{xy} \cdot \rho_{yz}$$

This equation can be equivalently written as:

$$\rho_{xz.y} = 0$$

since:

$$\rho_{xz.y} = (\rho_{xz} - \rho_{xy} \cdot \rho_{yz}) / [(1 - \rho^2_{xy}) \cdot (1 - \rho^2_{yz})]^{1/2}.$$

Such constraints on covariances are particularly interesting since it can be shown that in linear causal models their satisfaction depends *on the causal model alone*, that is, the dag, and not on the distribution of variables neither on particular numerical values of the coefficients a_{vu}. Thus, it is not necessary to estimate either the coefficients or the variances of the independent variables, since we can know which equations are satisfiable (or implied) by a SLCT $T = (\langle V,E\rangle, (W,P), X, L)$ by simply looking at the causal model $\langle V,E\rangle$. In fact, the following two theorems can be proven [4 pp. 265-288, 293-310]:

Theorem 1

Let $T = (\langle V,E\rangle, (W,P), X, L)$ be a SLCT, $\langle V,E\rangle$ an acyclic graph, and u, v, w, x four distinct measured variables in V. Then the following propositions are equivalent:

a) T implies the tetrad equation $\rho_{uv} \cdot \rho_{wx} = \rho_{ux} \cdot \rho_{vw}$

b) $\sum_{t \in T_{uv}} L(t) \cdot \sum_{t \in T_{wx}} L(t) \equiv \sum_{t \in T_{ux}} L(t) \cdot \sum_{t \in T_{vw}} L(t)$

where $\equiv$ means that the left expression identically equals the right expression, that is the two expressions are equal for all possible values of their linear coefficients.

For instance, the model in Figure 2 implies the following tetrad equation:

$$\rho_{xw} \cdot \rho_{zy} = \rho_{xz} \cdot \rho_{wy}$$

since the left expression of condition b) in theorem 1, that is:

$$(a_{w\lambda}a_{x\lambda} + a_{wz}a_{z\lambda}a_{x\lambda})(a_{y\lambda}a_{z\lambda} + a_{x\lambda}a_{yx}a_{z\lambda}) = a_{w\lambda}a_{x\lambda}a_{y\lambda}a_{z\lambda} + a_{w\lambda}a^2_{x\lambda}a_{yx}a_{z\lambda} + a_{wz}a_{x\lambda}a_{y\lambda}a^2_{z\lambda} + a_{wz}a^2_{x\lambda}a_{yx}a^2_{z\lambda}$$

identically equals the right expression, which is:

$$a_{x\lambda}a_{z\lambda}(a_{y\lambda}a_{w\lambda} + a_{w\lambda}a_{x\lambda}a_{yx} + a_{y\lambda}a_{z\lambda}a_{wz} + a_{x\lambda}a_{yx}a_{z\lambda}a_{wz})$$

Theorem 2

Let $T = (\langle V,E\rangle, (W,P), X, L)$ be a SLCT, $\langle V,E\rangle$ an acyclic graph, and x, y, z three distinct measured variables in V. Then the following propositions are equivalent:

a) $\rho_{xz.y} = 0$

b) y is a vertex in any trek connecting x to z, and either any trek between y and z is reduced to an acyclic path or any trek between y and x is reduced to an acyclic path. Formally: $(\forall\, t \in T_{xz} : y \in t) \wedge [(\forall\, t \in T_{yz} : t \in P_{yz}) \vee (\forall\, t \in T_{yx} : t \in P_{yz})]$.

For instance, the model in Figure 2 implies the following partial equation:

$$\rho_{vz.x} = 0$$

since x is a node of the only trek connecting v and z, $\langle\langle\lambda,x,v\rangle, \langle\lambda,z\rangle\rangle$, and the only trek connecting x and v is $\langle\langle x,v\rangle, \langle x\rangle\rangle$, which is an acyclic path.

In conclusion, the problem of checking whether a constraint (tetrad or partial equation) is satisfied by a SLCT is reduced to a search problem for some trek sets. Then the problem is moved to verifying constraints which hold in the data. In order to establish whether a tetrad equation holds, an asymptotic test is performed on the *tetrad difference*:

$$H_0 : \hat{\rho}_{uv} \cdot \hat{\rho}_{wx} - \hat{\rho}_{ux} \cdot \hat{\rho}_{vw} = 0$$

where $\hat{\rho}_{ij}$ is the sample correlation between the variables i and j. It can be proven that, as the sample size grows, the distribution of a tetrad difference converges in probability to

a normal distribution having mean zero and variance given by Wishart's formula [1].

As to the partial correlation, the following test:

$$H_0: \hat{\rho}_{xy.z} = 0$$

can be easily performed by taking into account the fact that the *Fisher's z* statistics:

$$z = \frac{1}{2} \ln \left(\frac{1+\hat{\rho}_{xy.z}}{1-\hat{\rho}_{xy.z}} \right)$$

converges in probability to a normal distribution with mean zero and variance given by $\sigma^2=1/(N-3)$, where N is the sample size [1].

When a certain causal model has been hypothesized, it is necessary to compare constraints which hold in the data to constraints implied by the model. Therefore, we need some *criteria* in order to evaluate how much a model fits to the data. Such criteria can also be exploited when comparing different alternative causal models. Glymour et al. [4] propose three criteria:

1) *H-I* (or *incompleteness* criterion): the number of constraints which hold in the data but are not implied by the model;
2) *I-H* (or *inconsistency* criterion): the number of constraints implied by the model but not holding in the data;
3) *simplicity* criterion: the previous criteria being equal, prefer the simplest model, that is the causal model with the lowest number of edges.

Obviously, the set *H* of constraints which hold in the data depends on the significance level used for the hypothesis tests presented above. Thus, when a significance level is fixed, the only way to change *H-I* and *I-H* is that of modifying the set *I* of implied constraints. In particular, the *specialization* of a causal model by adding an edge may decrease the number of implied constraints, in which case the model incompleteness (*H-I*) may increase while the model inconsistency (*I-H*) may decrease.

All these ideas have been implemented in a program, called TETRAD [4], that helps the user to search for good models of correlation or covariance data. The program operates in two modes: *manual* and *automatic*. In the former mode, the user can ask the program to provide information on the possible elaborations of a given causal model. For each elaboration, only one edge is added to the causal model, so that the user is given in charge of the task of making the best choice and then analyzing new possible elaborations of the modified causal model. On the contrary, TETRAD's automatic search procedure starts with an initial causal model and suggests the addition of *sets* of treks to the *initial* causal model in order to provide an *extended* model which better fits to the data according to the above criteria. The initial model must be an acyclic graph such that:

a) there are no edges between measured variables;
b) all measured variables are effect of a latent cause;
c) all latent variables are connected each other.

For each foursome of measured variables, a particular subgraph of the causal model is selected. Such subgraph consists of:

1) the foursome of measured variables;
2) all latent variables, named *parents*, which are adjacent to any variable in the foursome;

3) all edges from the parents to any variable in the foursome;
4) all treks, including latent variables which are not parents, between the parents.

According to the type of subgraph, TETRAD suggests the addition of sets of treks defeating the implication of constraints which do not hold in the data at a significance level α. However, some of the suggested treks may also increase the incompleteness of the model, in which case they will be excluded from further processing. The remaining suggested treks form *locally minimal* sets LM_i, since their addition to the model is evaluated by considering only a subgraph of the causal model. As search proceeds from foursome to foursome, it forms *globally minimal* sets GM_i of suggested treks such that GM_i will defeat as many implied constraints involving all foursomes considered so far as is possible without increasing incompleteness. At the end of the automatic search, TETRAD outputs the sets GM_i and the minimum significance level at which these sets are non-empty. Even if it is not explicitly stated in [4], tetrad equations are the only constraints considered by TETRAD when it builds the sets GM_i.

TETRAD's automatic procedure is heuristic since the addition of treks to the initial model is based only on local information, that is the type of subgraph built for each foursome. Thus, if the initial model satisfies the conditions a-c) listed above, the sets of suggested treks are probably, but not certainly, correct. However, if one of those conditions is violated, the sets of suggested trek additions are not even probably correct.

4 The Method of Conditional Independencies

This method builds the causal model explaining sample data by analysing the conditional independence which holds in the sample. It is based on a clear theory of causality which classifies several types of causal relationships. This is done in the light of Reichenbach's principle [10] according to which every dependence between two variables x and y has a causal explanation, so either one causes the other or there is at least a third variable z influencing both of them (z is the source of a trek between x and y). Moreover, also in this approach, temporal information is not considered essential, but it can be easily exploited when available.

Let O be a set of random variables with a joint probability distribution P, and $x, y, z_1, z_2, ..., z_n \in O$. Then x and y are said to be (*conditionally*) *independent in the context* $\{z_1, z_2, ..., z_n\}$ if it happens that:

$$P(x,y \mid z_1, ...,z_n) = P(x \mid z_1, ...,z_n) \cdot P(y \mid z_1, ...,z_n)$$

Henceforth, $I(P)$ will denote the set of conditional independencies between the variables in O when P is the joint probability distribution. Moreover, we will write $I(x,y/\{z_1,z_2,...,z_n\})$ if x and y are conditionally independent in the context $\{z_1, z_2, ..., z_n\}$, otherwise we will write $\neg I(x,y/\{z_1, z_2, ..., z_n\})$.

Given a causal model $M=<V,E>$ and its subset of observed variables $O \subseteq V$, we define *latent structure* the pair $S=<M,O>$. In other terms, S is a structure explaining which variables are observed in a causal model M.

Definition 5 ***(implication of a probability distribution)***
A latent structure $S=<M,O>$ *implies* a joint probability distribution P^* on O iff there exists a SCT $T=(<V,E>,(\Omega,P),X)$ such that $P_{[O]}=P^*$, where $P_{[O]}$ denotes the marginal distribution for the variables in O.

Henceforth, the set of probability distributions implied by a latent structure S will be denoted by P_S.

Definition 6 ***(consistency with a probability distribution)***

A latent structure $S=<M,O>$ is *consistent* with a given probability distribution P^* on the variables in O iff $P^* \in P_S$.

For a given sample distribution P^* on a set O of observed variables, we could try to find a latent structure $S=<M,O>$ that is consistent with P^*. However, the problem of estimating P^* is very hard, especially in the case of a large number of observed variables, even when the form of the probability distribution is known a priori. In order to simplify the problem, Pearl and Verma [7] reduce it to finding a latent structure that is consistent with the set of conditional independencies of P^*, $I(P^*)$, rather than P^* itself.

Definition 7 ***(implication of a conditional independence)***

A latent structure $S=<M,O>$ *implies* a conditional independence I between some variables in O, if for each $P \in P_S$ we have $I \in I(P)$, that is I is a conditional independence for each probability distribution implied by the latent structure.

Henceforth, the set of conditional independencies implied by a latent structure S will be denoted by $I(S)$. Formally, we can write:

$$I(S) = \bigcap_{P \in P_S} I(P)$$

The set $I(S)$ can be derived by simply looking at the causal model. Indeed, the following theorem can be proven:

Theorem 3

Given a latent structure $S=<M,O>$, and two variables $x,y \in S$, then there is a conditional independence between two variables x and y in a given context C iff

1. $\forall\, t \in T_{xy}\ \exists\, z \in C : z \in t$, and
2. $\forall z \in C$ $[(\forall t \in T_{zx}$ such that t does not contain variables in $C-\{z\} : t \in P_{zx}) \vee$
   $(\forall t \in T_{zy}$ such that t does not contain variables in $C-\{z\} : t \in P_{zy})]$

Figure 3a shows a latent structure and the corresponding set of conditional independencies which could be derived according to Theorem 3. It is worthwhile to observe that such a set of conditional independencies can be observed for the latent structure S' in Figure 3b, as well. However, this is possible for only some, not all, probability distributions with which the latent structure is consistent. Therefore, $I(x,y/\varnothing)$ is implied by S and not S' according to definition 7.

Definition 8 ***(consistency with a set of conditional independencies)***

A latent structure $S=<M,O>$ is consistent with a given set $I(P^*)$ of conditional independencies between the variables in O iff $I(P^*)=I(S)$.

As stated above, Pearl and Verma move the goal to finding a latent structure $S=<M,O>$ such that $I(S)=I(P^*)$. However, there is no guarantee that independencies observed on the sample distribution P^* are not *accidental* but they actually reflect the true (*structural*) independencies in the underlying causal model M. For this reason the two authors assume that P^* is a *stable distribution* generated by the underlying SCT.

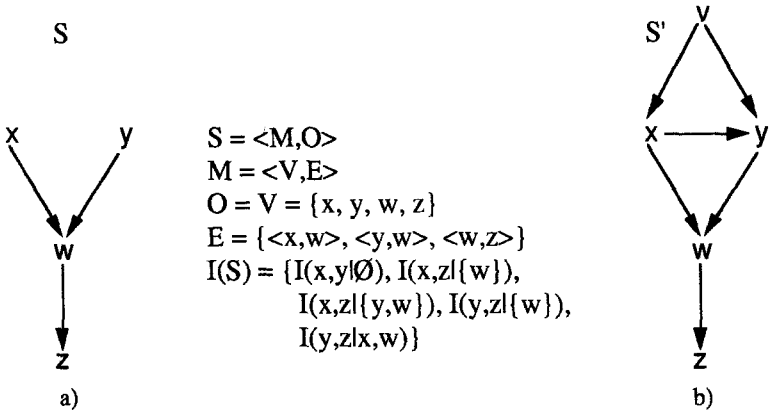

Figure 3. a) Latent structure and corresponding set of implied conditional independencies. The conditional independence I(x,y|∅) can be observed for some, but not all, particular probability distributions with which the latent structure S' in b) is consistent. Thus, it is not implied by S'.

Definition 9 ***(stable distribution)***

Let $S=<M,O>$ be a latent structure. A SCT $T=(M,(W,P),X)$ generates a *stable distribution* $P_{[O]}$ on O iff $I(P_{[O]}) \subseteq I(S)$, that is $P_{[O]}$ does not contain extraneous independencies.

Since by definition of $I(S)$ we have that $I(S) \subseteq I(P_{[O]})$, then for a stable distribution we can conclude that $I(S)=I(P_{[O]})$. Therefore, under the assumption of stability it is reasonable to restrict the search of a causal model to the latent structures that are consistent with $I(P^*)$. In this way we have assumed that Nature does not show a sample distribution P^* which implies accidental independencies but It can still hide some variables. Thus, the problem of finding a causal model is still under-constrained, since there could be an infinite number of dependency-equivalent latent structures $S=<M,O>$ *with* different numbers of latent variables and such that $I(S) = I(P_{[O]})$. Pearl and Verma restrict their search to particular latent structures called *projections*.

Definition 10 ***(projections)***

A latent structure $S=<<V,E>,O>$ is a *projection* on O of another latent structure $S'=<<V',E'>,O>$ iff

1) for each latent variable $\lambda \in V$ we have:
   a) indegree(λ)=0, that is λ is independent;
   b) outdegree(λ)=2;
   c) $\lambda \rightarrow v$, $\lambda \rightarrow w$ are in E, where $v,w \in O$ and v, w are not connected by any edge in E';
2) I(S)=I(S').

Therefore, the maximum number of latent variables which can be added to a graph is given by the number of pairs of observed variables not connected by any edge. An example is shown in Figure 4. Pearl and Verma [7] have proven that each latent structure has at least one projection, therefore it is reasonable to search a causal model only in the space of projections. For projections it is convenient to use hybrid graphs as representation formalism.

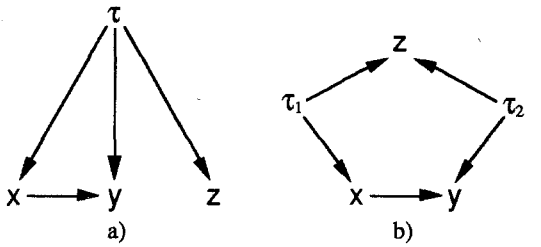

Figure 4. The latent structure in b) is a projection of the latent structure in a).

Definition 11 ***(hybrid graph)***

An *hybrid graph* is a couple $G=<V,E>$ where V is the set of nodes and $E=<E_N, E_U, E_B>$ is a triple of three disjunct sets of edges:

E_N is a set of *non-directed* edges (u—v);
E_U is a set of *uni-directed* edges (u→v);
E_B is a set of *bi-directed* edges (u↔v).

By extension, for each pair $u,v \in V$ we write that the edge $<u,v> \in E$ iff $(u—v) \in E_N$ or $(u \rightarrow v) \in E_U$ or $(u \leftrightarrow v) \in E_B$. According to the above definition, a directed graph is a particular hybrid graph in which $E_N = E_B = \varnothing$. Projections are represented by means of hybrid graphs in which $E_N = \varnothing$, moreover, a latent cause λ for two observed variables u and v is represented by eliminating the variable λ and adding a bi-directed edge between u and v. Thus the graph of a projection will have only observed variables. The intersection of the hybrid graphs representing all the possible projections of latent structures consistent with the sample distribution is called *core* of P^*.

The IC-algorithm (Inductive Causation algorithm) proposed by Pearl and Verma [7] builds the core of a sample distribution P^*, that is the intersection of the hybrid graphs representing all the possible projections of the minimal latent structure consistent with P^*. The input to the program is the set of conditional independencies which hold in P^*. The output core has four kinds of edges:

a) *marked uni-directed edges* representing genuine causal relationships;
b) *unmarked uni-directed edges* representing potential causal relationships;
c) *bi-directed edges* representing spurious associations;
d) *non-directed edges* representing those causal relationships that cannot be classified by using the only independencies in O (there would need different constraints).

Below, the definitions of genuine and potential causal relationships as well as the definition of spurious association are provided.

Definition 12 ***(potential causal influence)***

A variable x has a *potential causal influence* on another variable y if

1) x and y are dependent in every context (for each context C : $\neg I(x,y|C)$);
2) there exists a variable z and a context C such that:
   a) $I(x, z \mid C)$;
   b) $\neg I(z, y \mid C)$.

If there are no latent variables, the only cases satisfying definition 12 are those shown

in Figure 5a. Indeed, if it were y←x, then there would be a dependence between x and z, in contrast with a). But what can guarantee that x and y have no latent common cause, that is, x↔y ? If we were sure that all the variables involved in the phenomenon under study had been taken into account, condition 1) would be enough to avoid this eventuality. Nevertheless, since we do not have this certainty we can only postulate a *potential* cause of x on y.

Definition 13 ***(genuine causal influence)***

A variable x has a *genuine causal influence* on y if there exists another variable z such that:
1) x and y are dependent in every context and there is a context C such that:
   a) z has a potential causal influence on x;
   b) $\neg I(z,y \mid C)$;
   c) $I(z,y \mid C \cup \{x\})$

or

2) x and y are in the transitive closure of rule 1.

Figure 5b shows the only two cases satisfying condition 1) when there are only three measured variables. Condition 2) covers those cases in which there is a path from x to y whose edges represent genuine causal relationships oriented in the direction from x to y.

Definition 14 ***(spurious association)***

Two variables x and y have a *spurious association* if they are dependent in some context S ($\neg I(x,y \mid C)$) and there exist two variables z_1 and z_2 such that:
1) $\neg I(z_1, x \mid C)$;
2) $I(z_1, y \mid C)$;
3) $\neg I(z_2, y \mid C)$;
4) $I(z_2, x \mid C)$.

In this definition, conditions 1) and 2) prevent x from causing y, while conditions 3) and 4) prevent y from causing x, thus the dependence between x and y can only be explained by a spurious association. An example of spurious association is shown in Figure 5c. In this case the context C is the empty set.

5 Empirical Comparison of the Two Methods

In this section, Glymour et al.'s and Pearl and Verma's approaches to causal inference are empirically compared. The first three experiments have been performed by using models in which the assumptions of linearity, normality, acyclicity and stability are made. Note

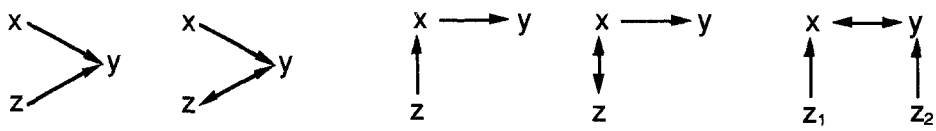

Figure 5. a) the two cases of potential causal influence; b) the two cases of genuine causal influence; c) a case of spurious association.

that when independent variables of a linear model are normal, then all random variables in the model are normal as well as the joint probability distribution. Thus, tests on independence between two variables, say x and y, in the context $\{z_1, z_2, ..., z_n\}$ are reduced to tests of Fisher's z statistics computed for the correlation coefficient $\rho_{xy.z_1, z_2, ..., z_n}$.

Obviously when all the variables involved in the phenomenon are observed, then the method of conditional independencies is by far the best. Indeed, since the sample distribution is stable, there is a causal connection between two variables in the graph only when they are dependent in each context. It could happen that some of these connections are not explained (not directed edges) or not precisely defined (potential causes) due to the lack of further control variables. However, when there are no latent variables and we know it, we can clearly claim that all potential causes are genuine causes, since there cannot be any spurious association without latent variables.

Experiment 1

Let us consider the model in Figure 6a. All independent variables have a standard normal distribution, N(0,1), the significance level for all the test is α=0.05 and the sample size is N=1000. In Table I, all the sample correlations are reported. From these correlations we find the following conditional independencies:

$$I(x,w \mid \varnothing) \qquad I(x, z \mid \varnothing) \qquad I(y, z \mid \{w\})$$

In Figure 6b, the output of the IC-algorithm is shown. Initially, the connections x—y, y—w and w—z are recovered, since the pairs of variables *(x,y)*, *(y,w)* and *(w,z)* are dependent in every context. Then, by applying twice the definition of potential causal inference at each pair of variables, the connections x→y and w→y are recovered, for x can be a control variable for the pair *(y,w)* and w can be taken as a control variable for the pair *(x,y)*. The connection w—z cannot be explained due to the lack of control variables. The connections x→y, w→y labelled as potential causes can be considered genuine causes if we assume a priori that there are no latent variables.

By trying to analyse this sample with TETRAD's automatic search procedure, we soon meet a problem: what is the initial model? In Figure 6c the simplest initial model is presented. Such a model is evidently wrong since the latent variable τ does not exist in reality. However, we will ignore connections with τ, since we are interested in finding causal relationships between observed variables alone.

The automatic search procedure suggests an addition of either a trek between x and y or a trek between w and z at the significance level α=0. Indeed, the only tetrad equation that holds in the data is:

$$\rho_{xw} \cdot \rho_{yz} = \rho_{xz} \cdot \rho_{yw}$$

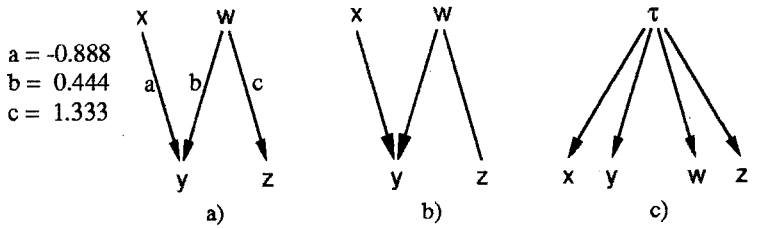

Figure 6. a) True causal model of experiment 1. b) Causal model inferred by means of the IC-algorithm. Bold arrows represent potential causes. c) TETRAD's initial causal model.

Table I

	x	y	w	z
x	1			
y	-0.6299	1		
w	-0.05351	0.3253	1	
z	-0.0566	0.2872	0.7930	1

Table II

	x	y	w	z
x	1			
y	0.5322	1		
w	0.4741	0.8659	1	
z	0.0057	0.0182	0.2471	1

while the other two tetrad equations implied by the model, that is:

$$\rho_{xy} \cdot \rho_{wz} = \rho_{xw} \cdot \rho_{yz}$$
$$\rho_{xy} \cdot \rho_{wz} = \rho_{xz} \cdot \rho_{yw}$$

do not. The addition of the a of either a trek between *x* and *y* or a trek between *w* and *z* improves the consistency criterion, but when we add the edges x→y and w→z at the initial model we get a model in which no further improvement is possible. To sum up, the method of constraints on correlations is not able to discover the causal relation w→y.

Experiment 2

When there are latent variables, the goodness of final results strongly depends on the topology of the initial causal model. For instance, let us consider the true causal model in Figure 7a, where τ is a latent variable. Once again the a sample is generated by imposing all independent variables to have a standard normal distribution. Table II shows the correlations between observed variables for a sample of 1000 observations. According to such correlations the following independencies are found:

$I(x,w \mid \{y\})$ $I(x,z \mid \varnothing)$ $I(y,z \mid \varnothing)$

By applying the IC-algorithm, the model in Figure 7b is built. This result is undoubtedly better than that obtained by TETRAD's automatic search procedure when the initial causal model is that shown in Figure 7c. Once again, this model includes a latent variable τ that does not exist in reality, but it is the simplest initial model we can provide. Since we are interested in finding causal relationships between observed variables we will ignore connections with τ. TETRAD suggests the addition of either a trek between *x* and *y* or a trek between *z* and *y*, even though there is no trek between *z* and *y* in the original model (*z* and *y* are independent in any context). Such treks could be realized by adding the following connections: x↔y and z→y. However, the final model is not better than the initial one, neither can we discover the causal relationship y→w.

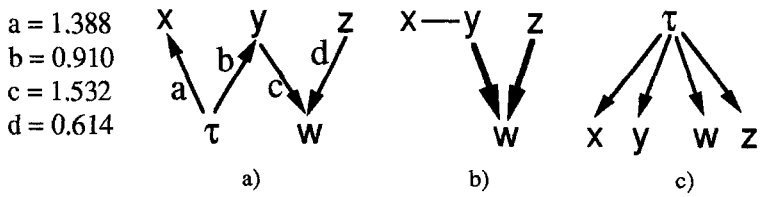

Figure 7. a) True causal model of experiment 2. b) Causal model inferred by means of the IC-algorithm. Bold arrows represent potential causes. c) TETRAD's initial causal model.

a = 0.865
b = 0.456
c = 1.217
d = 0.398
e = 0.567
f = 0.854
g = 0.772
h = 0.593
i = 0.626

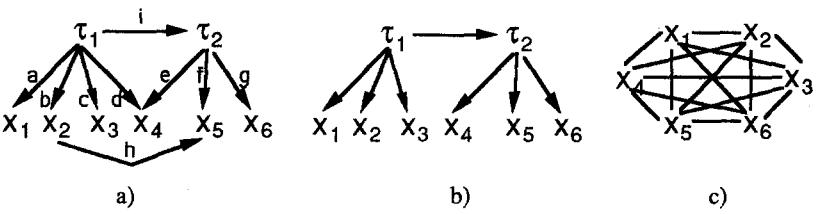

Figure 8. a) True causal model of experiment 3. b) TETRAD's initial causal model. c) Causal model inferred by means of the IC-algorithm.

Experiment 3

In this experiment we generated a sample of 5000 observations for the model in Figure 8a and we ran TETRAD with the initial model of Figure 8b. Correlations are given in Table III. In this case a trek between x_2 and x_5 is suggested and by asking a detailed analysis of the initial model augmented with the edge $x_2 \rightarrow x_5$ we are provided with further suggestions, namely:

$x_4 \rightarrow x_5$ $\quad x_4 \leftrightarrow x_5$ $\quad \tau_1 \rightarrow x_4$ $\quad x_6 \rightarrow x_5$ $\quad \tau_1 \rightarrow x_6$ $\quad x_6 \leftrightarrow x_5$

By testing all these additions to the augmented model, we find that the best models are the following:

initial model + $x_2 \rightarrow x_5$ + $\tau_1 \rightarrow x_4$
initial model + $x_2 \rightarrow x_5$ + $x_6 \rightarrow x_5$
initial model + $x_2 \rightarrow x_5$ + $x_6 \leftrightarrow x_5$

among which there is the true causal model.

The method of conditional independencies provides poor results since no independencies hold in the data. In fact, it is possible to discover independencies only between observed variables, but having excluded latent variables from consideration all couples of variables are determined to be dependent (see Figure 8c).

To sum up, the method of conditional independencies is not good to study those phenomena in which there are some latent variables controlling many observed variables which have no spurious association. Indeed, this method can only discover projections of the real model, where a latent variable can only control two observed variables.

Table III

	x1	x2	x3	x4	x5	x6
x1	1					
x2	0.25120	1				
x3	0.49724	0.30191	1			
x4	0.36036	0.20448	0.42635	1		
x5	0.32524	0.51905	0.38348	0.48400	1	
x6	0.23871	0.12966	0.27900	0.44256	0.47144	1

Experiment 4

In this experiment we generated a sample of 1000 observations concerning the model in Figure 9. Each independent observed variable has an ordinal domain with three possible values: low (L), medium (M) and high (H). In this case there is no linear dependence between the variables, even if the causal relationships are still probabilistic. Of course, conditional independencies cannot be tested by means of the Fisher's z statistics as in the previous experiments, since the variables are not integer or ratio level measurements. We built a contingency table for each pair of variables in order to test the independence in the empty context, and we used a χ^2 test on each table. For testing conditional independence for non-empty contexts we had to generate as many subtables as the number of possible values that variables in the context can take. At the significance level α=0.05 we detected the following independencies:

$I(x_1, x_2 \mid \varnothing)$ $I(x_1, x_4 \mid \varnothing)$ $I(x_1, x_5 \mid \{x_3\})$ $I(x_2, x_4 \mid \varnothing)$ $I(x_2, x_5 \mid \{x_3\})$ $I(x_3, x_4 \mid \varnothing)$

The IC-Algorithm builds the original model with the only difference that all the causal relationships but $x_3 \rightarrow x_5$ are potential and not genuine.

At this point, having discovered the causal dependencies between the variables, we used a learning system in order to induce the causal rules from the data. In particular, two learning problems were defined: the first for learning the causal law which relates the study level and the intelligence of a student with his/her preparation, and the second for learning how the preparation and the leniency of the examiner can affect the final result of the examination. In order to solve both problems we used C4.5 [9], a learning system that induces decision trees from examples. In particular, in the first problem examples are described by means of x_1, x_2 and the class (target) attribute x_3, while in the second problem examples are described by means of x_3, x_4 and the class (target) attribute x_5. Results are shown in Figure 10. The composition of the training cases at a leaf F gives a probability $P(K / F)$ that a case at F belongs to class K, where the probability can be estimated as a relative frequency [8]. For instance, in the case of a moderately lenient examiner and an insufficiently prepared student, the probability of having a low score is 71/(71+16) = 0.82 while the probability of having a medium score is 16/(71+16) = 0.18. From tree b) in

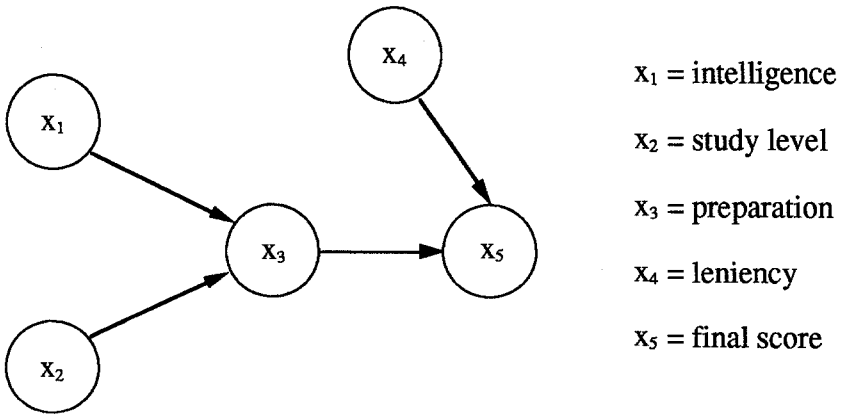

Figure 9. Causal model of experiment 4. All variables have an ordinal domain.

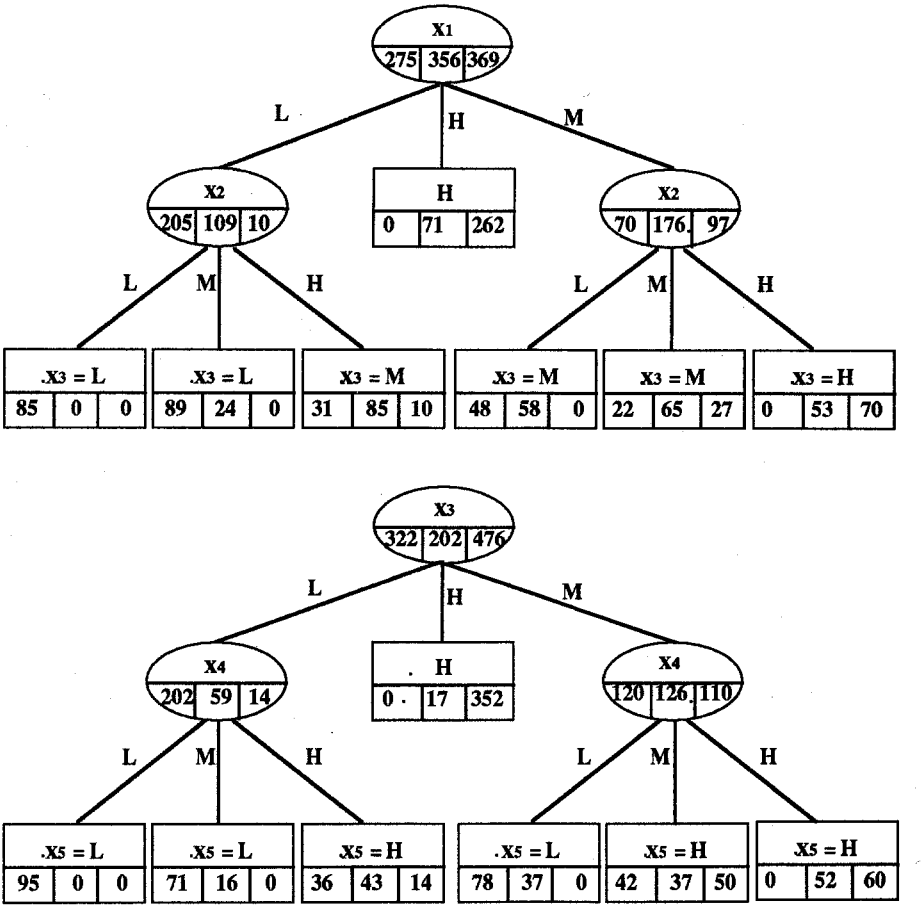

Figure 10. The two decision trees generated by C4.5 for the model in Figure 9. Together, they explain the causal law that rules the sociological model. Leaves are represented in boxes. Triplets of numbers in each node represent the distribution of examples reaching the node with respect to the values taken by the target attribute, namely L, M, and H.

Figure 10, we can also draw the conclusion that even students with an insufficient or moderately good preparation can get a high score when the examiner is particularly lenient. These and other rules can be directly derived in an explicit form by transforming each decision tree into a set of production rules.

6 Conclusions

Inferring causal models from empirical data is an arduous but exciting task. In fact, causal knowledge is a relevant part for any intelligent system that aims at explaining its decisions, predictions or even the behaviour of another intelligent system [2]. However, most of the causal relationships that humans use in their reasoning are probabilistic.

In this paper an analytic and empirical comparison between two different approaches to statistical causal inference has been presented. These approaches differ both in the kind

of constraints they consider while selecting a causal model and in the way they search for the model which better fits to the sample data.

In Glymour et al.'s approach, constraints on correlation implied by the model are compared with those that hold in the data. Two kinds of constraints are considered: tetrad and partial equations, involving foursomes and triplets of variables respectively. TETRAD is a program that helps to discover causal models by testing tetrad differences and partial correlations. It can work in two different modes: manual and automatic. In the automatic mode a heuristic search strategy is performed in order to find an elaboration of the initial causal model that improves the completeness and consistency criteria with respect to tetradic constraints. The automatic procedure is heuristic since the addition of treks to the initial model is based only on local information, that is, the type of subgraph built for each foursome. TETRAD simply helps to discover causal models but it neither performs any statistical test of the causal model as a whole nor it estimates the coefficients of the model.

However, the true problem with TETRAD is just in the asymptotic test of tetrad differences and partial correlations. As Glymour et al. themselves admit, tests are performed on each tetrad difference or partial correlation as though it were independent of the others, even if this is almost never correct. In fact, there is no well-known test for the whole set of tetrad differences or partial correlations involving the very same variables.

Note that the strong assumptions underlying Glymour et al.'s approach are linearity and normality. The former limits the applicability to datasets containing only interval or ratio level measurements. Another limitation is the necessity of disposing of a complete initial model that involves latent variables.

These limitations do not concern the approach based on conditional independencies which in turn suffers from the very opposite problem of not exploiting partial initial knowledge of the causal structure. Pearl and Verma's approach exploits constraints determined by conditional independencies in order to detect genuine causal influences among observed variables. Note that, under the assumptions of normality and linearity, a test on the partial equation $\rho_{xz.y} = 0$ is equivalent to a test of the conditional independence $I(x, z / \{y\})$. Thus, this approach exploits some of the constraints used in the other one. Nevertheless, tetrad equations are constraints peculiar of Glymour et al.'s approach, while conditional independencies with contexts having a cardinality different from one are constraints exploited only in Pearl and Verma's approach.

There are a bias and a basic assumption in the method of conditional independencies: the former is towards projections while the latter concerns the stability of the underlying distribution. As already shown in the third experiment, such a bias can prevent the IC-algorithm from discovering the correct underlying model. On the contrary, stability seems a reasonable assumption that helps to reduce the search space. The applicability of Pearl and Verma's approach to variables of any level of measurement is simply limited by the availability of statistical tests for independence between variables of different levels of measurement.

Some preliminary experiments show that the two methods are complementary. Nevertheless, they share some common aspects. Firstly, both approaches require a large sample of data. Indeed, tests on tetrad differences are positively biased for small samples while several tests on conditional independencies may lead to unpredictable results if we use high significance levels. Secondly, it is possible to cast the search performed by both

methods as a constraint satisfaction problem, where constraints are determined by the approaches themselves. This last aspect will be investigated in future work.

Finally, we have pointed out that both methods investigated in the paper can only discover causal dependencies between variables in the model but they are not able to define the causal laws. In the fourth experiment we have shown how, using an inductive learning algorithm, it is possible to fill this gap. In particular, we generated probabilistic decision trees which can effectively represent probabilistic causal relationships.

Acknowledgments

Thanks to Michael Pazzani for his helpful comments on an earlier draft of the paper and to Francesco Colasuonno for his precious collaboration on conducting the experiments. Also, a special thank to the anonymous reviewers whose comments helped to make this paper clearer.

References

1. T. W. Anderson: An introduction to multivariate statistical analysis. New York: Wiley 1958
2. P.R. Cohen, A. Carlson, L. Ballesteros, R. St. Amant: Automating path analysis for building causal models from data. In Machine Learning: Proceedings of the Tenth International Conference. San Mateo: Morgan Kaufmann 1993, pp. 57-64
3. G. DeJong, R. Mooney: Explanation-based learning: an alternative view. Machine Learning 1, 145-176 (1986)
4. C. Glymour, R. Scheines, P. Spirtes, K. Kelly. Discovering causal structure. Orlando: Academic Press 1987
5. D. Heise. Causal analysis. New York: Wiley 1975.
6. T.M. Mitchell, R.M. Keller, S.T. Kedar-Cabelli: Explanation-based generalization: a unifying view, Machine Learning 1, 47-80 (1986).
7. J. Pearl, T. S. Verma: A theory of inferred causation. In: J. A. Allen, R. Fikes, E. Sandewall (eds.): Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference. San Mateo: Morgan Kaufmann 1991, 441-452
8. J. R. Quinlan: Probabilistic Decision Trees. In: Y. Kodratoff, R.S. Michalski (eds): Machine learning: an artificial intelligence approach, vol III. San Mateo: Morgan Kaufmann 1990, 140-152
9. J. R. Quinlan: C4.5: programs for machine learning. San Mateo: Morgan Kaufmann 1993
10. H. Reichenbach: The direction of time. Berkley: University of California Press 1956
11. P. Spirtes, C. Glymour, R. Scheines: Causation, prediction and search. Berlin: Springer 1993
12. L. Steels: Second Generation Expert Systems. Future Generation Computer Systems 1, 213-221 (1985)
13. P. Suppes: A probabilistic theory of causation. Amsterdam: North Holland 1970