The Representation and Learning of Equivalent Information in Causal Models TECHNICAL REPORT IRIS-TR-0099, May 2006

Jan Lemeire
ETRO Dept.
Vrije Universiteit Brussel
Brussels, Belgium
jan.lemeire@vub.ac.be

Erik Dirkx
ETRO Dept.
Vrije Universiteit Brussel
Brussels, Belgium
erik.dirkx@vub.ac.be

Sam Maes
COMO lab
Vrije Universiteit Brussel
Brussels, Belgium
sam.maes@vub.ac.be

Stijn Meganck
COMO lab
Vrije Universiteit Brussel
Brussels, Belgium
stijn.meganck@vub.ac.be

Abstract

From causal theory it is known that the independencies entailed by deterministic relations in a stochastic distribution cannot be represented by a faithful causal model. Deterministic relations lead to situations in which either of two variables X and Y become conditionally independent from a third variable Z by conditioning on the other variable. More generally, this occurs when X and Y contain the same information about Z, they are called *information-equivalent*. The joint distribution defines an *equivalent partitioning* of the domains of X and Y by which only the states are related for which the conditional distribution of target Z is the same, hence $P(Z \mid X) = P(Z \mid Y)$. We propose to select the relation with the target variable containing the least complexity. Under the assumption that complexity does not increase along a Markov chain, this selection criterion results in consistent models. Faithfulness of the graph can be reestablished by limiting the conditional independencies by the simplicity criterion in cases of equivalent information. On the other hand, all conditional independencies among the variables can be retrieved from the graph by a generalized definition of the d-separation property. Finally, the PC algorithm was extended to learn models containing information-equivalent variables from data.

1 INTRODUCTION

A causal model offers a representation of the conditional independencies ($\perp\!\!\!\perp$) of a joint distribution by a Directed Acyclic Graph (DAG). The graphical condition of *d-separation* ($\perp$) allows to determine the dependency of two variables. A causal model is called *faithful* if there is a complete correspondence between $\perp\!\!\!\perp$ and $\perp$ - the model provides a dense description of the relational regularities found in the data. To develop the theory of causal models, Pearl used an axiomatic characterization of probabilistic dependency [Pea88]. Four properties of conditional independencies between variables are taken as axioms: symmetry, decomposition, weak union and contraction. A fifth condition, intersection, only holds for strictly positive distributions:

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W}, \boldsymbol{Y} \wedge \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W}, \boldsymbol{X} \Rightarrow \boldsymbol{X}, \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W} \quad (1)$$

where $A \perp\!\!\!\perp B \mid C$ stands for the independency of A and B by conditioning on C. The condition claims that if 2 variables depend on Z, it is impossible that each of the two variables will render the other irrelevant. Most theorems of causal theory do not depend on the intersection condition, but some of the important theorems, such as the uniqueness of minimal I-maps, are based on it. Faithfulness is the main motivation for the causal interpretation of Bayesian networks. The independendencies in data are qualitative regularities and if a minimal, unique model is able to predict all regulurities, it must come close to reality.

A well-known case for which the intersection condition is not valid, is when the model contains deterministic relations. Most research in causal models therefore restricts itself to probabilistic relations. The researchers argue that variables that are completely determined by others are not essential in the model since they contain redundant information. Our approach is developed in the perspective of building information models for offering insight into computer systems. Performance models contain many deterministic relations. But even when application and system parameters determine the performance completely, intermediate variables that influence the outcome should be included in the model for providing insight in the results. Examples are memory usage, cache misses or number of floating point operations. Fig. 1 shows a performance model of a sort algorithm. At the left are the parameters, such as the size of the array to be sorted, the type of the elements, the processor clock frequency, etc. Through intermediate variables, like the number of basic operations $\#operations$ and the time of one operation T_{1op}, the computation time T_{comp} is calculated.

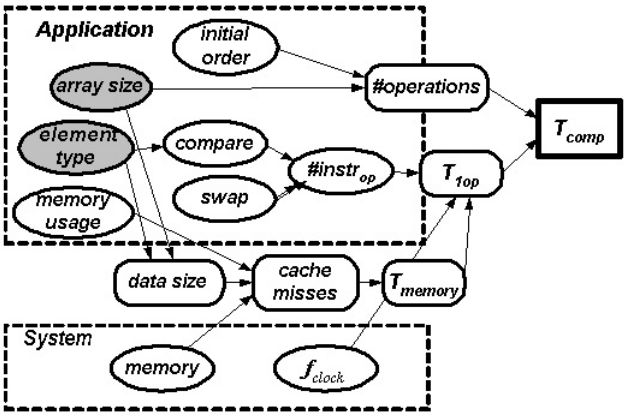

Figure 1: Performance model of a sort application with parameters $arraysize$ and $elementtype$

Recent research developed methods for performing inferences in Bayesian Networks with functional dependencies [CDLS99, CS04]. Geiger and Spirtes et all. developed a condition, *D-separation*, to detect the dependencies entailed by deterministic relations from the causal model [Gei90, SGS93]. Our approach intends to integrate information-equivalence into causal models such that faithfulness is guaranteed. We will show that the problematic case in which

$$\boldsymbol{X \perp\!\!\!\perp Z \mid Y \ \wedge \ Y \perp\!\!\!\perp Z \mid X \ \Rightarrow X, Y \perp\!\!\!\perp Z} \qquad (2)$$

is violated, happens when X and Y contain *equivalent information* about Z, a more general condition than just a deterministic relation. Current causal modeling does not support the faithful representation of such situations. The left-hand conditional independencies suggest that the edges $X - Z$ and $Y - Z$ should not be present in the graph. But if X or Y are correlated with Z, at least one of both should be adjacent to Z in the absence of other variables. The solution here proposed also holds under conditioning, as expressed by Eq. (1). Conditioning can be regarded as a new joint distribution constructed by constraining the variables of the conditioning set W to certain states.

The next section recalls some basic properties of conditional independencies and section 3 gives an introduction to causal models. Section 4 discusses the related work. Section 5 defines exactly when nodes contain equivalent information about others. In section 6, the criterion of complexity is introduced to identify adjacency among equivalent variables. Then, section 7 shows how the faithfulness of causal models can be reestablished. The soundness and correctness is proven in section 8 and finally, the PC algorithm is adapted to handle cases of information-equivalence.

2 PRELIMINARIES

Two stochastic variables X and Y are **conditionally independent** by conditioning on Z if $P(Y \mid X, Z) = P(Y \mid Z)$. Independency can also be interpreted in information-theoretic terms. Two variables contain information about each other if they are dependent - by knowing one variable, the uncertainty - or entropy - of the other is reduced. Dependency of variables can be quantified by the reduction in uncertainty, called **mutual information**, written as $I(X;Y)$. The standard definitions can be found in [CT91].

Definition 1 *Random variables X, Y, Z are said to form a Markov chain in that order (denoted by $X \Rightarrow Y \Rightarrow Z$) if the joint probability mass function can be written as*

$$P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y) \qquad (3)$$

Important consequences are as follows [CT91]:

- X, Y, Z form a Markov Chain if and only if X and Z are conditionally independent given Y.
- Information cannot increase along a Markov chain: $I(X;Y) > I(X;Z)$. This is called the *Data Processing Inequality*.
- $X \Rightarrow Y \Rightarrow Z$ implies that $Z \Rightarrow Y \Rightarrow X$ is also a Markov chain. Therefore, the condition is sometimes written $Z \Leftrightarrow Y \Leftrightarrow X$.

If we consider that X contains information about Z in the Markov chain $X \rightarrow Y \rightarrow Z$, the Markov chain is present in the causal models $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$ or $X \leftarrow Y \rightarrow Z$, but not in the *v-structure* $X \rightarrow Y \leftarrow Z$. Here, the arrows have a causal interpretation. The last of the four models is responsible for the asymmetry of causality. The variable Y is called a *collider* along the path from X to Z. We will write a Markov chain for which X and Z are dependent, written as $X \not\perp\!\!\!\perp Z$, as $X - Y - Z$.

A causal model is represented by a directed acyclic graph (DAG), in which adjacent nodes represent a direct causal relation between both variables.

If two variables X and Y are dependent, implied by $P(Y \mid X) \neq P(Y)$, the conditional distribution of one variable differs for at least two values of the conditioning variable: $\exists x_1, x_2 \in X_{domain} : P(Y \mid x_1) \neq P(Y \mid x_2)$. The information a variable contains about another lies into the differences in the conditional distributions. Values for which this distribution is the same contain the same information.

Definition 2 *The domain of X, denoted by X_{dom}, can be partitioned into disjoint subsets X^k_{dom} for which $P(Y \mid x)$ is the same for all $x \in X^k_{dom}$. We call this the $Y - partition$ of X_{dom}. We define $\kappa_Y(X)$ as the index of the subset.*

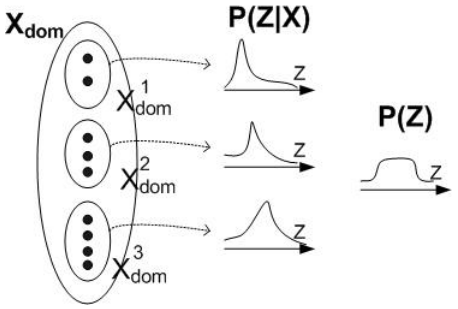

Figure 2: Y-partition of the domain of X

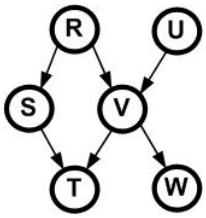

Figure 3: Example causal model.

Accordingly, the conditional distribution depends solely on the index of the partition:

$$P(Y \mid X) = P(Y \mid \kappa_Y(X)) \tag{4}$$

Fig. 2 shows a Y-partition of X_{dom} and the related conditional distributions of Y.

3 CAUSAL MODELS

Causal models are graphical models that explicitly describe the relational properties of variables [Pea00] [SGS93] [TP02]. Causal models are also Bayesian networks, which offer dense representations of joint distributions. The Directed Acyclic Graph (DAG) of Fig. 3 corresponds to the following factorization:

$$P(R,S,T,U,V,W) = \\ P(R).P(S \mid R).P(T \mid S,V).P(U).P(V \mid U).P(W \mid V). \tag{5}$$

The causal interpretation of a Bayesian network is based on the *faithfulness* property: the model describes *all* relational properties that can be found in the data, which are characterized by the conditional independencies.

By an independency, the conditional distribution can be rewritten:

$$U \perp\!\!\!\perp W \mid V \Leftrightarrow P(U \mid W) = \sum_{v \in V} P(U \mid v).P(v \mid W) \tag{6}$$

All information U and W have about each other goes through V. This is a consequence of the model of Fig. 3.

The graphical d-separation criterion allows us to retrieve the conditional independencies from the graph that follow from the *Markov condition*. It states that a node becomes independent from all its non-descendants by conditioning on its parents.

Definition 3 *(d-separation) Let p be a path between a node U and a node V of a DAG G. Path p is called blocked given subset **W** of nodes in G if there is a node w on p satisfying one of the following conditions:*

1. *w has converging arrows (along p) and neither w nor any of its descendants are in **W**, or*
2. *w does not have converging arrows (along p) and w is in **W**.*

***W** is said to **d-separate** U from V in G, denoted $U \perp V \mid$ **W**, iff they block every path from U to V.*

In the model of Fig. 3, U and W get d-separated by V, R and T by S and V. R and U are d-separated, but are not d-separated if V is given. $R \rightarrow V \leftarrow U$ is called a *v-structure*. Conditioning unblocks a v-structure in a path whereas it blocks non-v-structures.

The faithfulness property insists that for all variables A, B and C a conditional independence in the distribution corresponds to a d-separation in the graph:

$$A \perp\!\!\!\perp B \mid C \Leftrightarrow A \perp B \mid C. \tag{7}$$

A causal model is a unique and minimal decomposition of the joint distribution into independent blocks: the conditional distributions $P(X_i \mid parents(X_i))$. The model offers a canonical description of the system under study, which is able to explain all observable regularities [Lem06]. This motivates, but can of course not ensure, that the model is a representation of the underlying physical mechanisms of the system.

Besides the formalization of causal networks, another breakthrough was the development of algorithms that can learn causal models from observational data [Pea00] [SGS93]. They are discussed in section 9.

4 RELATED WORK

Recent research developed methods for performing inferences in Bayesian Networks with functional dependencies [CDLS99, CS04]. Dechter and Mateescu introduce mixed networks for expressing probabilistic and deterministic information in the form of constraints [DM04], whereas we view deterministic relations as proper causal relations. Geiger, Spirtes et al. extended the d-separation criterion for retrieving the dependencies entailed by deterministic relations, which they called *D-separation* [Gei90, SGS93].

Learning algorithms require that functionally determined variables are eliminated from the input data. The argument is that such variables are not essential to the model since they contain redundant information. Determinism is, however, not always known a priori. Besides, such variables can provide insight in the underlying mechanisms and often reduce the complexity of the model. The idea is to augment the causal model with the knowledge of the deterministic relations. This makes it possible to reestablish faithfulness by introducing an additional criterion - the complexity of a relation - for choosing among equivalent relations.

Pearl uses *stability* as the main motivation for the faithfulness of causal models [Pea00]. Take the model of Fig. 3. In general, T will depend on R. They will be independent only when the stochastic parametrization is such that the influences via paths $R \rightarrow S \rightarrow T$ and $R \rightarrow V \rightarrow T$ cancel out exactly . This is unstable because a small change in the parametrization will result in a dependency. The unhappy balancing act is a measure zero event, the chance of such a coincidence can therefore be regarded as having zero probability. Deterministic relations, however, appear in nature and are not coincidences.

For the faithfulness of graphical models, a lot of conditions should hold [Pea88]. Therefore, other representation schemes of independency information were developed, such as the imsets of Studeny [Stu01], which can model any conditional independence structure. Our approach claims that if the violations of the conditions are based on local properties, they can be integrated into causal models.

5 VIOLATION OF THE INTERSECTION CONDITION

First is demonstrated how deterministic relations violate Eq. (2). Then, the notion of equivalent partition and information is introduced to capture all situations in which the intersection condition becomes untrue.

5.1 DETERMINISTIC RELATIONS

We will denote a set of variables $(Y_1, Y_2, \ldots, Y_n)$ by $\boldsymbol{Y}$ and an outcome of $\boldsymbol{Y}$ as $\boldsymbol{y}$.

Definition 4 *A variable X is **determined** by a set of variables $\boldsymbol{Y}$ if for every $\boldsymbol{y} \in \boldsymbol{Y}$: $P(X \mid \boldsymbol{Y}) = 1$ for exactly one value of X and zero for all other values. The relation between X and the set $\boldsymbol{Y}$ is a function, written as $X = f(\boldsymbol{Y})$.*

A functional relation implies that the variables $\boldsymbol{Y}$ contain all information about X. The uncertainty of X becomes 0 due to the knowledge of $\boldsymbol{Y}$.

Theorem 1 *If X is determined by $\boldsymbol{Y}$ then $\forall Z : X \perp\!\!\!\perp Z \mid \boldsymbol{Y}$*

Proof:
By the deterministic relation, for each value of $\boldsymbol{Y}$, X is completely known and thus $P(X \mid \boldsymbol{Y}) = 1$ for $X = f(\boldsymbol{Y})$ and 0 for the other $X - values$. Any conditional distribution can be rewritten as

$$P(Z \mid X, \boldsymbol{Y}) = \frac{P(Z, X \mid \boldsymbol{Y})}{P(X \mid \boldsymbol{Y})} \quad \forall X : P(X \mid \boldsymbol{Y}) > 0 \quad (8)$$

The denominator is only strictly positive for $X = f(\boldsymbol{Y})$, so

$$P(Z \mid X, \boldsymbol{Y}) = \frac{P(Z, f(\boldsymbol{Y}) \mid \boldsymbol{Y})}{P(f(\boldsymbol{Y}) \mid \boldsymbol{Y})} = \frac{P(Z \mid \boldsymbol{Y})}{1} \quad (9)$$

∎

If X and Z are correlated, all information from X about Z is also present in Y. If additionally $Y \perp\!\!\!\perp Z \mid X$ holds, meaning that Y contains no additional information about Z, the intersection condition is violated. For a bijection $X = f(Y)$, *all* variables dependent of X or Y imply that $X \perp\!\!\!\perp Z \mid Y$ and $Y \perp\!\!\!\perp Z \mid X$.

It can be argued that a deterministically related variable can safely be removed from the model, since the determining variables contain all its information and make the variable redundant. This is correct, but there are good reasons to keep them in the model. Consider the variable *data size* in the performance model of Fig. 1. It is determined by *element type* and *array size*. However, it is extremely useful in providing insight how *element type* and *array size* influence the *cache misses*. Secondly, it reduces the complexity of the model. Intermediate variables can be the known variables - which can be measured, for instance - or the target of interventions. And the next section will demonstrate that also probabilistic relations can lead to Eq. (2).

In the rest of the discussion, we will only consider single variables. The results can easily be extended to sets of variables. This can be done by replacing the set $(X_1, X_2, \ldots, X_n)$ by a single variable $\boldsymbol{X} = X_1 \times X_2 \times \ldots \times X_n$.

5.2 EQUIVALENT PARTITION

Definition 5 *A relation $\Re \subset X \times Y$ defines an **equivalent partition** in Y_{dom} to a partition of X_{dom} if:*

1. *$\forall x_1$ and $x_2 \in X_{dom}$ that do not belong to the same partition: $\forall y_1 \in Y_{dom}$ with $x_1 \Re y_1$, it must be that $\neg(x_2 \Re y_1)$.*
2. *For all subsets X^k_{dom} of the partition: $\exists x_1 \in X^k_{dom}, \exists y_1 \in Y_{dom} : x_1 \Re y_1$.*

Fig. 4 shows an example of an equivalent partitioning. No y is related to X-values belonging to different partitions

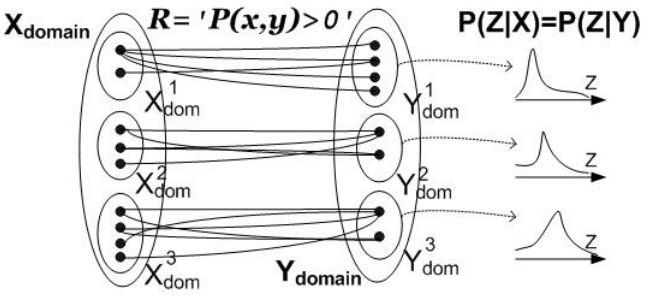

Figure 4: Relation $\Re$ defines an equivalent partition to X_{dom} in Y_{dom}

and for every partition, there is at least one related Y-value. Note that a function, for which every x-value has a related Y-value, defines an equivalent partition on Y_{dom} for every partition of X. The next theorem defines exactly when the intersection condition, as expressed by Eq. 2, is violated. We consider the relation $\Re$ defined by $P(x, y) > 0$, so that two values $x \in X_{dom}$ and $y \in Y_{dom}$ are related if there is a chance that both appear together.

Theorem 2 *If $X \not\perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z \mid X$, then*

$X \perp\!\!\!\perp Z \mid Y \Leftrightarrow$ *the relation $x \Re y$ defined by $P(x, y) > 0$, with $x \in X_{dom}$ and $y \in Y_{dom}$, defines an equivalent partition in Y_{dom} to the Z-partition of X_{dom} .*

Proof $\Leftarrow$
We have to prove that $P(Z \mid Y, X) = P(Z \mid Y)$. The left hand side leads to $P(Z \mid Y, X) = P(Z \mid X) = P(Z \mid \kappa_Z(X))$, with $\kappa_Z(X)$ the index of X in the Z-partition of X_{dom}. We will prove that $P(Z \mid Y)$ leads to the same expression.
By the independence $Y \perp\!\!\!\perp Z \mid X$, we can write that

$$P(Z \mid Y) = \sum_{x \in X_{dom}} P(Z \mid X = x).P(X = x \mid Y) \quad (10)$$

The last factor only differs from zero if both X and Y belong to the subsets that correspond to each other by the equivalent partition. By proper numbering of the subsets, the indices correspond. It follows that

$$= \sum_{x \in X_{dom}: \kappa_Z(x) = \kappa_Z(y)} P(Z \mid X = x).P(X = x \mid Y) \quad (11)$$

$$= P(Z \mid \kappa_Z(x)) \sum_{x \in X_{dom}: \kappa_Z(x) = \kappa_Z(y)} P(X = x \mid Y) \quad (12)$$

Since the conditional distribution of Z is constant in each subset of the Z-partitioning by definition. The sum is 1, since $P(X = x \mid Y)$ is zero everywhere else.

Proof $\Rightarrow$
We have to show that $\forall x_1, x_2 \in X_{dom}$ for which $P(Z \mid x_1) \neq P(Z \mid x_2)$, there exists a $y_1 \in Y_{dom}$ for which $P(x_1, y_1) > 0$ and that for all such y_1 values $P(x_2, y_1) = 0$.
Since $P(x_1) > 0$, there must be at least one value y_1 for which $P(x_1, y_1) > 0$, otherwise $P(X, Y)$ is not a valid distribution. Next, both given conditional independencies, $Y \perp\!\!\!\perp Z \mid X$ and $X \perp\!\!\!\perp Z \mid Y$, imply that

$$P(Z \mid x_1, y_1) = P(Z \mid x_1) = P(Z \mid y_1). \quad (13)$$

Assume that $P(x_2, y_1) \neq 0$, then likewise

$$P(Z \mid x_2, y_1) = P(Z \mid x_2) = P(Z \mid y_1) \quad (14)$$

Combining the right hand sides of both equations leads to the contradiction that $P(Z \mid x_2) = P(Z \mid x_1)$. ∎

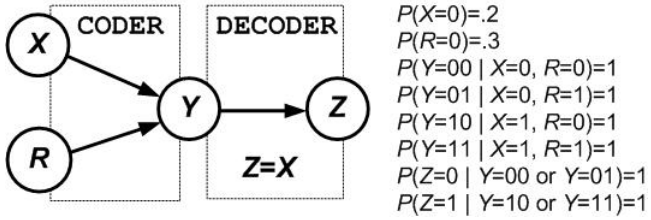

Figure 5: Causal model of SGS in which Z equals X(Fig. 3.23)

It follows that $I(Y; Z) = I(X; Z)$, hence X and Y contain the same *amount* and the same *kind* of information about Z.

5.3 EQUIVALENT INFORMATION

Definition 6 *X and Y contain* ***equivalent information*** *about Z, if*

- *$X \not\perp\!\!\!\perp Z$ and $Y \not\perp\!\!\!\perp Z$*
- *$Y \perp\!\!\!\perp Z \mid X$*
- *$X \perp\!\!\!\perp Z \mid Y$*

X and Y are called ***equivalent variables*** *with respect to Z.*

Take the coder-decoder example, taken from SGS (Fig. 3.23) [SGS93], shown by Fig. 5. Variable Y encodes the values of both R and X, and Z decodes Y to match the value of X. X is therefore deterministically related to Z, though not adjacent. X is related to Z through Y. The conditional distributions reveal the $Z - partition$ of Y_{dom}, $\{\{0, 1\}, \{2, 3\}\}$, and the distribution defines an equivalent partition in X: subset $\{0, 1\}$ of Y_{dom} corresponds with subset $\{0\}$ of X_{dom} and $\{2, 3\}$ with $\{1\}$. Both X and Y contain equivalent information about Z.

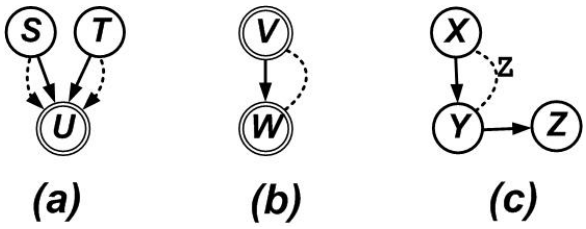

Figure 6: Augmented Causal Model with a Functional Relation (a), a Bijection (b) and an Information-Equivalence (c).

By the equivalence of X and Y for Z it follows that

$$P(Z \mid X) = P(Z \mid Y). \tag{15}$$

Knowledge of either X or Y is completely equivalent with respect to Z. By the information-equivalence the Z-partitions of X_{dom} and Y_{dom} are related one-to-one by the equivalent partition. Each subset of one partition corresponds to one subset of the other partition. By proper renumbering of the indices of both partitions, one can write that

$$P(Z \mid \kappa_Z(X)) = P(Z \mid \kappa_Z(Y)) \tag{16}$$

5.4 NOTATION

We propose the following notation. Deterministic nodes are depicted with double-bordered circles with dashed edges coming from the determining variables, as shown in Fig. 6 (a). If the parents comprise all the determining variables, the dashed edges may be omitted. Two variables related by a bijection are linked with an unoriented dashed edge (Fig. 6 b). Information-equivalent variables are connected by a dashed edge annotated with the target variable (Fig. 6 c).

5.5 QUASI-DETERMINISTIC RELATIONS

Any stochastic model can be emulated by functional relationships of the form

$$x_i = f_i(pa_i, u_i), \;\; i = 1, \ldots, n \tag{17}$$

where pa_i stands for the set of variables judged to be immediate causes of X_i and where U_i represent errors due to omitted factors [Pea00]. The random disturbances U_i make the relations probabilistic, they ensure that adjacent nodes share exclusive information, and they are therefore the key for being able to recognize direct relations. Practically, the errors have to be sufficiently large and enough data should be available to be able to identify adjacency in experimental data. If not, indirect relations observationally resemble direct ones and situations of information-equivalence appear.

6 THE COMPLEXITY CRITERION

Consider two variables X and Y information-equivalent with respect to Z. If either of both gets independent from Z by conditioning on a certain subset, the other would also. Consequently, none of the variables should be adjacent to the target in the model. In the absence of such subset, the equivalent variables contain exclusive information about the target. This should be reflected by the model, but poses a problem for a faithful graphical representation of the distribution. The two independencies, $Y \perp\!\!\!\perp Z \mid X$ and $X \perp\!\!\!\perp Z \mid Y$, suggest that neither X or Y are adjacent to Z, but this would fail to represent the exclusive information both variables contain about the target. Including both edges disrupts the minimality condition, since both variables have the same information about the target.

Consider the two possible cases for X and Y having equivalent information about Z:

1. $X \not\perp\!\!\!\perp Y \mid Z$. X and Y will be related through a path not containing Z. Relating one of both with Z suffices to model the information they contain about Z.

2. $X \perp\!\!\!\perp Y \mid Z$. All three variables contain equivalent information about each other. Two of the three possible edges to connect them are necessary to reflect the dependencies. In the causal model of Fig. 5, X, Y and Z reflect this case.

6.1 THE COMPLEXITY OF RELATIONS

We need criteria, different from the conditional independencies, to select among information-equivalent relations. Such criteria could depend on properties of the variables or the relations among them. The characteristic of the variables relevant for the target variable Z is the information they contain about Z; which is reflected by the uncertainty about the Z-partition. Because of equivalence, this uncertainty is the same, $P(\kappa_Z(X)) = P(\kappa_Z(Y))$, such that no characteristic provide a valid selection criterion. The only objective criterion available is the complexity of the relations, according to which - in the spirit of Occam's Razor - simpler relations should be preferred over complex ones. Accordingly, the choice between two equivalent variables X and Y for being adjacent to the target node Z is deciding upon which relation - $P(Z \mid X, \boldsymbol{U})$ or $P(Z \mid Y, \boldsymbol{U})$ - is the simplest, where the set $\boldsymbol{U}$ comprises the other parent nodes of Z.

In practice, the feasibility of the complexity quantification of the relations cannot be guaranteed, but becomes more plausible by the *reductionist* paradigm of causality. A fundamental property of causal models is that they break up into independent, local submodels, in which the relations $P(node \mid parents)$ represent basic mechanisms lying close to the physical mechanisms. They will be relatively

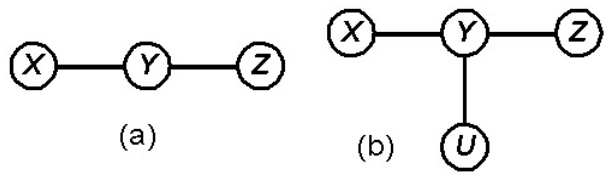

Figure 7: Complexity increase in a Markov Chain (a) and with a fourth variable (b).

simple compared to the complexity of the overall model. For continuous variables, a regression analysis can recognize the type of the analytical function $x_i = f(pa_i, u_i)$, which determines the relation's complexity. For discrete variables, the conditional distributions are described by a discrete distribution. The number of probabilities in the probability table determine the complexity, unless local regularities allow further compression of the distribution [BFGK96]. Methods of objectively quantifying complexity are discussed in [LV97, GMP05].

When the complexities of the relations match, we advocate to let the choice undecided and keep both edges. Knowledge of the information-equivalence will make it possible to use the model appropriately. If, for example, X and Y are related by a linear bijection, the relation with any other variable is similar. There are no objective criteria for determining adjacency. There is a rationale for this, since both variables are indistinguishable in the perspective of the system under study. They contain equivalent information about any other variable, so - in the absence of background knowledge - they represent equivalent quantities.

6.2 COMPLEXITY INCREASE

The complexity criterion makes sense by the validity of the following assumption:

Assumption 1 *The Complexity Increase assumption:*

1. *If $X \not\perp\!\!\!\perp Z$ and $X \perp\!\!\!\perp Z \mid Y$, described by Fig. 7(a), then the relation between X and Z is not less complex than the relations $X - Y$ and $Y - Z$.*

2. *If the relation $X - Y$ is more complex than $Y - Z$, and if $X \not\perp\!\!\!\perp U$ and $X \perp\!\!\!\perp U \mid Y$, then the relation $X - U$ is more complex than $Z - U$. See Fig. 7(b).*

The assumption states that the complexity between variables do not decrease along a causal path. This would happen by correspondences of the relations and neutralization of its complexities, which can be regarded as yet another regularity. This additional regularity should also be added to the model, but this falls out of the scope of this paper. Note the similarity of the assumption with the *Data Processing Inequality*, discussed in section 2.

Reconsider the model of Fig. 5, in which X and Y are equivalent for Z. The relation $X \rightarrow Z$ is simpler than $Y \rightarrow Z$. The complexity increase assumption is violated, due to a complete dependence of the decoding relation $Y \rightarrow Z$ on both $X \rightarrow Y$ and $R \rightarrow Y$. However, it is completely 'natural' that an inference algorithm considers the relation $X - Z$ as a direct one and not the more complex $Y - Z$. The learned model will not correspond to the real model, but we claim that it is not possible to learn the 'true' model from observation alone. Note that the simpler model is appropriate for prediction of Z, but would fail to predict the intervention $do(Y)$.

7 FAITHFULNESS

For capturing information-equivalence in a faithful causal model, we should reconsider the properties we want to be shown graphically and the condition for retrieving independencies from the graph.

7.1 CONDITIONAL INDEPENDENCE AND SIMPLICITY

To ensure the correspondence between the graph and the distribution, the definition of conditional independency has to be generalized with the requirement that the conditioning set should provide a simpler relation if containing equivalent information.

Definition 7 ***Conditional independence and simplicity*** *between two variables X, Y and a conditioning set Z is true when*

- $X \perp\!\!\!\perp Y \mid Z$
- *And the relation $Z - Y$ is less complex than $X - Y$ if $Z \perp\!\!\!\perp Y \mid X$ (Z and X contain equivalent information about Y).*
- *And the relation $Z - X$ is less complex than $Y - X$ if $X \perp\!\!\!\perp Z \mid Y$ (Z and Y contain equivalent information about X).*

We write this as $X \perp\!\!\!\perp_S Y \mid Z$.

This definition reestablishes the Markov condition for a causal model, according to which the parents of a node cannot become independent by conditioning on some subset, while non-descendants become independent by conditioning on the parents.

7.2 D-SEPARATION

When there are deterministic relationships among variables, there are conditional independencies that are not entailed by the Markov condition alone. SGS [SGS93], based

on the work of Geiger [Gei90], enlarged the concept of d-separation to create a graphical condition for retrieving all conditional independencies from a graph and a set of deterministic relations.

The extended D-separation includes the independencies that follow from deterministic relations.

Definition 8 *(D-separation) Let p be a path between a node X and a node Y of a DAG G. Path p is called blocked given subset **Z** of nodes in G if there is a node w on p satisfying one of the following conditions:*

1. *w has converging arrows (along p) and neither w nor any of its descendants are in **Z**, or*
2. *w does not have converging arrows (along p) and w is in **Z** or is determined by **Z**.*

Z *and the set of deterministic relations is said to D-separate X from Y in G, denoted $X \perp Y \mid$ **Z**, iff they block every path from X to Y.*

More generally, variables can contain equivalent information about X or Y with variables along the path between X and Y. These cases also entail additional conditional independencies. The definition of D-separation is generalized to capture these independencies.

Definition 9 *(D_{eq}-separation) Let p be a path between a node X and a node Y of a DAG G. Path p is called blocked given subset **Z** of nodes in G and a set of information equivalences if there is a node w on p satisfying one of the following conditions:*

1. *w has converging arrows (along p) and neither w nor any of its descendants are in **Z**, or*
2. *w does not have converging arrows (along p) and w is in **Z** or has an equivalent variable for a node along the path in **Z**.*

Z *and the set of deterministic relations is said to D_{eq}-separate X from Y in G, denoted $X \perp Y \mid$ **Z**, iff they block every path from X to Y.*

Take the Markov chain of Fig. 8, from the original definition of *d-separation*, A and B become separated by X, but not by Y. If, however, Y is information-equivalent to X with respect to Z, A and B also become independent by conditioning on Y. Consider that

$$P(B \mid X) = P(B \mid Z)P(Z \mid X)$$

$$= P(B \mid Z)P(Z \mid Y) = P(B \mid Y). \quad (18)$$

The dependency of B on X is similar to that of Y.

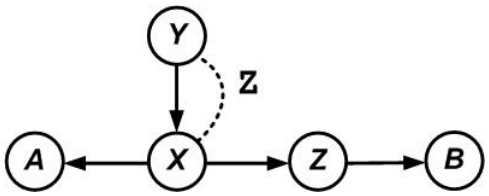

Figure 8: $D_{eq} - separation$ in a model with nodes X and Y having equivalent information about Z

7.3 FAITHFULNESS REDEFINITION

Given the additional independencies that equivalence relations entail, the definition of faithfulness should be reconsidered. Since the situation in which the independencies $X \not\perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z \mid Y$ and $Y \perp\!\!\!\perp Z \mid X$ hold cannot be represented graphically, we restricted the conditional independencies by the definition of conditional independency and simplicity $\perp\!\!\!\perp_s$. On the other hand, equivalence of nodes with respect to a target node, affects the conditional independencies from the target node with other nodes. These independencies can be found graphically with the *D_{eq}-separation* condition $\perp_{eq}$.

Definition 10 *A causal model is called faithful to a probability distribution containing variables with equivalent information if*

$$X \perp_{eq} Y \mid Z \;\Leftrightarrow\; X \perp\!\!\!\perp Y \mid Z \quad (19)$$

$$X \perp Y \mid Z \;\Leftrightarrow\; X \perp\!\!\!\perp_s Y \mid Z \quad (20)$$

8 SOUNDNESS AND COMPLETENESS

In this section we will prove that the redefinition of faithfulness is sound and complete, ie. all possible consequences of equivalent information are captured by the DAG. Take the equivalence case depicted in Fig. 9. We postulate that $Y \rightarrow X$ carries the same information as $X \rightarrow Z$. But what if other variables got involved in this path. What if the influence from X on Z goes through another variable or Z is related to other variables. How will the information-equivalence of X and Y propagate and will the equivalence of edge $Y \rightarrow X$ with $X \rightarrow Z$ correct with respect to other variables. Fig. 10 shows 4 possibilities of other variables getting involved. This section investigates under which assumptions the notation, which has an intuitive interpretation, is correct, namely it can be read of from the graph that:

- X and Y are equivalent for A and C.
- B gets conditionally independent from Z by Y (and D).
- D is also equivalent with respect to Z.

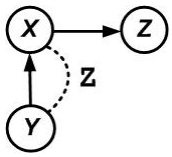

Figure 9: X and Y are information-equivalent for Z.

8.1 Assumptions

Two assumptions are made for the study of the effect of information-equivalences on other relations. The first is that an information-equivalence remains when conditioned on other variables:

Assumption 2 *Information-equivalence remains under conditioning*

$$\boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{X} \;\&\; \boldsymbol{X} \perp\!\!\!\perp Z \mid \boldsymbol{Y} \;\Leftrightarrow\; \boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{X, W} \;\&\; \boldsymbol{X} \perp\!\!\!\perp Z \mid \boldsymbol{Y, W} \tag{21}$$

for all subsets of variables $\boldsymbol{W}$*, disjoint from* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$*, and not containing* Z*.*

This is certainly true for deterministically related variables. The second assumption is *weak transitivity* [Pea88].

Assumption 3 *Weak Transitivity*

$$\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{W} \;\&\; \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{W, U} \;\Rightarrow\; \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{U} \mid \boldsymbol{W} \; \textit{or} \; \boldsymbol{U} \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{W} \tag{22}$$

It is one of the necessary conditions for the existence of a faithful graph. The condition expresses a form of transitivity. If T depends on U and U depends on V, it must that either T depends on V (eg. model $T \rightarrow U \rightarrow V$) or becomes dependent by conditioning on U (*v-structure* $T \rightarrow U \leftarrow V$)

Take again the coder-decoder example shown in Fig. 5. We see that X affects the first bit of Y and R the second. The decoding of Z should thus be determined by the first bit of Y. This model however violates the weak transitivity condition. Y depends on X, R and Z, but R is independent from X and Z, also after conditioning on Y. The condition demands that if two variables are connected through a chain of variables in which the adjacent variables are dependent, they should also be dependent (under conditioning).

8.2 Properties of Equivalent Information

First, we prove 4 properties about the effect of an information-equivalence on other variables. Fig. 10 represents a model containing the 4 cases.

Property 1 *If X and Y are information-equivalent with respect to a variable Z, $X \not\perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp A \mid Z$, then*

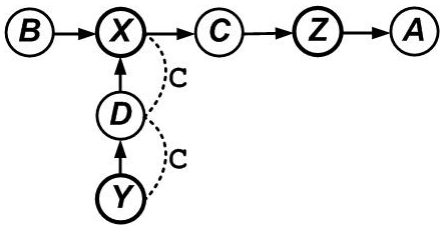

Figure 10: Model with X and Y equivalent for Z. The other nodes show possible consequences.

- $Y \perp\!\!\!\perp A \mid Z$
- *X and Y are information-equivalent with respect to A*

Proof:
Besides the given $X \perp\!\!\!\perp A \mid Z$, X stays independent from A when conditioning on Z and Y. $X \not\perp\!\!\!\perp A \mid Z, Y$ would mean that X is connected to A with a path not containing Z, but containing Y in a v-structure [1]. But then Y would be connected to Z with a path via A not containing X, implying $Y \not\perp\!\!\!\perp Z \mid X$ or $Y \not\perp\!\!\!\perp Z \mid X, A$, which both contradict the given information-equivalence.

Then, by applying weak transitivity on $X \perp\!\!\!\perp A \mid Z$ and $X \perp\!\!\!\perp A \mid Z, Y$, it follows that $X \perp\!\!\!\perp Y \mid Z$ or $Y \perp\!\!\!\perp A \mid Z$ should hold. The first independence is not true, so the second should hold which proves that Y is also independent from A given Z.

Information-equivalence of X and Y with respect to A follows from

$$\begin{aligned} P(A \mid X) &= \sum_{z \in Z} P(A \mid z).P(z \mid X) \\ &= \sum_{z \in Z} P(A \mid z).P(z \mid Y) = P(A \mid Y) \end{aligned} \tag{23}$$

The last step is true by $Y \perp\!\!\!\perp A \mid Z$ (cf Eq. 6). ∎
This case is shown in Fig. 10 by node A.

Property 2 *If X and Y contain equivalent information about a variable Z, which also holds under conditioning, it follows that*

$$Z \perp\!\!\!\perp B \mid X \Leftrightarrow Z \perp\!\!\!\perp B \mid Y \tag{24}$$

Proof:
By assumption 2, it follows that

$$P(Z \mid \boldsymbol{B}, \boldsymbol{Y}, \boldsymbol{W}) = P(Z \mid \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{W}) \tag{25}$$

$$= P(Z \mid \boldsymbol{X}, \boldsymbol{W}) = P(Z \mid \boldsymbol{Y}, \boldsymbol{W}) \tag{26}$$

[1] This follows from the Markov condition as given by the d-separation criterion, which reflects all consequences of Markov (section 3).

Node B in Fig. 10 illustrates this case.

Property 3 *If X and Y contain equivalent information about a variable Z, it follows that*

$$X \perp\!\!\!\perp Z \mid C \Leftrightarrow Y \perp\!\!\!\perp Z \mid C \tag{27}$$

Proof:

$$P(Z \mid C, Y) = P(Z \mid C, X) = P(Z \mid C) \tag{28}$$

For the equivalence, we have to show that $Y \perp\!\!\!\perp C \mid X$ and $X \perp\!\!\!\perp C \mid Y$. From $Y \perp\!\!\!\perp Z \mid X$ and $Y \perp\!\!\!\perp Z \mid C, X$ (by Eq. 21), weak transitivity demands that $Y \perp\!\!\!\perp C \mid X$ or $C \perp\!\!\!\perp Z \mid X$. The second independence is not true, which proofs the first independence. The independence $X \perp\!\!\!\perp C \mid Y$ is proved with the same arguments. ∎
This is shown in Fig. 10 by node C.

Property 4 *If X and Y contain equivalent information about a variable Z, it follows that*

$$X \perp\!\!\!\perp Y \mid D \Rightarrow X, Y \perp\!\!\!\perp Z \mid D \tag{29}$$

Proof:

$$\begin{aligned} & X \perp\!\!\!\perp Z \mid Y \\ \Rightarrow\ & P(X \mid Z) = P(X \mid Y).P(Y \mid Z) \\ & \qquad = P(X \mid D).P(D \mid Y).P(Y \mid Z) \\ & \qquad = P(X \mid D).P(D \mid Z) \end{aligned} \tag{30}$$

D has all information that X and Y share, so it also must have all information they have about Z. ∎

8.3 Soundness and Completeness

We prove the soundness and completeness of Conditional and Simplicity Independence and D$_{eq}$-separation by showing that the consequences of combinations of information-equivalence and other conditional independencies are captured consistently by the model, ie. that they do not lead to unfaithful situations. Take X and Y being Z-equivalent and the relation $X - Z$ simpler than $Y - Z$. This is expressed by the properties $Y \perp\!\!\!\perp_S Z \mid X$, $X \perp\!\!\!\perp Z \mid Y$ and $X \not\perp\!\!\!\perp_S Z \mid Y$. By the refined definition of faithfulness, the properties are represented faithfully by model $Y - X - Z$. X and Y are connected and also X with Z. The conditional and simplicity independence of Y and Z by X, makes that X lies on the path connecting Y and Z.

We have to investigate the implications for other conditional independencies found in the distributions. An independence statement with one variable of the three variables X, Y or Z involved in it only says something about a path with that variable. There are nine possible combinations of using 2 of the three variables in a conditional independence statement.

1. $X \not\perp\!\!\!\perp A$ and $X \perp\!\!\!\perp A \mid Z$:
   By property 1 it follows that $Y \perp\!\!\!\perp A \mid Z$ and that X are Y equivalent for A. The second part of the Complexity Increase Assumption assures that $Complexity(X, A) < Complexity(Y, A)$, thus $Y \perp\!\!\!\perp_S A \mid X$, but $X \not\perp\!\!\!\perp_S A \mid Y$. This is shown in Fig. 10.

2. $Z \not\perp\!\!\!\perp B$ and $Z \perp\!\!\!\perp B \mid X$:
   The independence $Z \perp\!\!\!\perp B \mid Y$ follows from property 2. Then, there are two possibilities:

   - If B has less information about Z, it is related to Z via X as shown in Fig. 10. By D_{eq}-separation the conditional independence $Z \perp\!\!\!\perp B \mid Y$ can be retrieved from the graph.
   - If on the contrary variable B contains as much information about Z as X, all three nodes are equivalent for Z. This is shown by node D in Fig. 10. The node having the simplest relation with Z is related to Z, which is X in the figure.

3. $Z \not\perp\!\!\!\perp C$ and $X \perp\!\!\!\perp Z \mid C$:
   By property 3, Y also gets independent, $Y \perp\!\!\!\perp Z \mid C$. There are two possible cases:

   - If $C \perp\!\!\!\perp Z \mid X$, C is also information-equivalent with respect to Z, which is discussed in the previous case.
   - If $C \not\perp\!\!\!\perp Z \mid X$, C has more information about Z. Property 3 proves that X or Y are information-equivalent for C as well, shown in Fig. 10. By the second part of the Complexity Increase Assumption, $X - C$ must be simpler than $Y - C$, since $X - Z$ must be simpler than $Y - Z$ and $X \perp\!\!\!\perp Z \mid C$.

4. $X \perp\!\!\!\perp Y \mid D$:
   By property 4, given below, it follows that $X \perp\!\!\!\perp Z \mid D$, which is discussed by case 3.

5. The last combination, $X \perp\!\!\!\perp E \mid Y$, only interferes with the Z-equivalence of X and Y if there is an independence with Z. This is discussed by the previous cases.

The 5 remaining cases, $Y \perp\!\!\!\perp A \mid Z$, $Z \perp\!\!\!\perp B \mid Y$, $X \perp\!\!\!\perp Z \mid C$, $Y \perp\!\!\!\perp Z \mid D$ and $Y \perp\!\!\!\perp D \mid X$, are equivalent to respectively cases 1, 2, 3, 4 and 5.

9 LEARNING ALGORITHM

One of the basic causal structure learning algorithms, which learns a model from data, is the PC algorithm, developed by Spirtes, Glymour and Scheines [SGS93]. It is proven that it constructs the correct graph for distributions that are faithful to some directed acyclic graph. It is of the constraint-based type, as opposed to scoring-based algorithms. Consult [KN03] for an overview. The graph is constructed in two steps. The first step, called fast-adjacency procedure, learns the undirected graph and the second tries to orient the edges. The construction of the undirected graph is based on the property that two nodes are adjacent if only if they remain dependent by conditioning on every set of nodes that does not include both nodes. The algorithm starts with a complete undirected graph and removes edges for each independence that is found. The number of nodes in the conditioning set is gradually increased upto a certain number, called the *depth* of the search. The orientation step is based on the recognition of the v-structure $X \rightarrow Y \leftarrow Z$, for which X and Z are independent, but become dependent conditional on Y. The result of the algorithm will be a set of models that are observationally indistinguishable. Two DAGs are proven to be observationally equivalent if and only if they have the same undirected graph and the same sets of *v-structures* [VP91]. It cannot be guaranteed that all edges can be oriented.

We now discuss how the PC algorithm should be adapted in order to learn the generalized models.

9.1 Equivalence Detection

Information-equivalence poses a problem for the algorithm. Take X and Y equivalent for Z, by $Y \perp\!\!\!\perp Z \mid X$ the algorithm would remove the $Y - Z$ edge and $X \perp\!\!\!\perp Z \mid Y$ deletes the $X - Z$ edge. Information-equivalences should therefore be detected during the construction of the undirected graph. For each conditional independence that is found, it should be tested whether an equivalence can be found by swapping the conditioning set with one of both arguments. For independence $U \perp\!\!\!\perp V \mid W$, the independence $W \perp\!\!\!\perp V \mid U$ would mean that U and W are equivalent for V, while $U \perp\!\!\!\perp W \mid V$ implies that V and W are equivalent for U. If both independencies are found, it means that the three variables are equivalent. We call this a *3-node-equivalence*. Each information-equivalence is recorded and does not lead to an edge removal.

The fast-adjacency procedure starts testing dependencies without conditioning, then conditions on one node, followed by conditioning on two nodes and so on. Since the edges involved in information-equivalences are not removed, conditional independence will be checked again in the next phase when extra variables are added to the conditioning set. For $Y \perp\!\!\!\perp Z \mid X$, it follows that $Y \perp\!\!\!\perp Z \mid X, U$ holds for any variable U. These independencies are consequences of the same equivalence, so these tests should be skipped in the procedure.

9.2 Complexity Calculation

Before starting with the orientation step of the algorithm, the complexity of the relations is used as criterion to select among equivalent edges. The complexity of all equivalent edges is estimated and a choice is made if a significant discrepancy is found. If the estimation fails or results in an insignificant difference, equivalent edges may remain in the model. We have shown that faithfulness is not endangered by it. For all *n-nodes-equivalences*, the $n - 1$ simplest edges should remain in the model such that all nodes are connected.

9.3 Orientation

The original orientatien rules can be applied on the undirected graph containing information-equivalences. Node Y has to be regarded as separated from Z by X, while X is not separated from Z by Y. The three variables involved in an information-equivalence form a Markov Chain $Y - X - Z$. This is confirmed by the conditional and simplicity independencies that follow from the information-equivalence. They forbid the v-structure $Y \rightarrow X \leftarrow Z$.

10 CONCLUSIONS

The intersection condition, on which causal theory is based, is not only broken by deterministic or quasi-deterministic relations, but more generally, when two variables containing *equivalent information* about a third variable. Only if the stochastic distribution generates an *equivalent partition* of both variables' domain, the information about the third is completely transferred from one variable to the other. Information-equivalence is a regularity that interferes with the conditional independencies that a causal model intends to describe correctly. To attain minimality and faithfulness, the presence of these regularities should be incorporated into the model. This paper proposes the complexity of the relations as criterion to determine adjacency among information-equivalent relations. Faithfulness can then be reestablished by enlarging the definition of conditional independency with the requirement of simplicity in cases of information-equivalence. On the other hand, the conditional independencies that are generated by nodes containing equivalent information can be retrieved from the graph by the $D_{eq} - separation$ property. The complexity criterion leads to consistent models by assuming that the complexity of the relations increases for more distant variables. Violation of the Complexity Increase Assumption is caused by yet another regularity - the correspondence of relations along a causal path.

The soundness and completeness of the augmented model was proven. It showed that we can speak about information flowing through the graph.

References

[BFGK96] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

[CDLS99] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

[CS04] Barry R. Cobb and Prakash P. Shenoy. Inference in hybrid bayesian networks with deterministic variables. In *in P. Lucas (ed.), Proceedings of the Second European Workshop on Probabilistic Graphical Models (PGM-04)*, pages 57–64, 2004.

[CT91] Thomas M. Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[DM04] Rina Dechter and Robert Mateescu. Mixtures of deterministic-probabilistic networks and their and/or search space. In *AUAI '04: Proc. of the 20th conf. on Uncertainty in artificial intelligence*, pages 120–129, Arlington, Virginia, United States, 2004. AUAI Press.

[Gei90] D. Geiger. *Graphoids: A Qualitative Framework for Probabilistic Inference*. PhD thesis, University of California, Los Angeles, 1990.

[GMP05] P. Grünwald, I.J. Myung, and M.A. Pitt. *A Tutorial Introduction to the Minimum Description Length Principle*. MIT Press, 2005.

[KN03] Kevin B. Korb and Ann E Nicholson. *Bayesian Artificial Intelligence*. CRC Press, 2003.

[Lem06] Jan Lemeire. Causal models as minimal descriptions of multivariate systems. http://parallel.vub.ac.be/∼jan, 2006.

[LV97] Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, Morgan Kaufman Publishers, 1988.

[Pea00] J. Pearl. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[SGS93] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 2nd edition, 1993.

[Stu01] Milan Studeny. On non-graphical description of models of conditional independence structure. In *HSSS Workshop on Stochastic Systems for Individual Behaviours*, Louvain la Neuve, Belgium, January 2001.

[TP02] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *AAAI/IAAI*, pages 567–573, 2002.

[VP91] T. Verma and J Pearl. Equivalence and synthesis of causal models. In *In Proc. of the 6th workshop on uncertainty in Artifical Intelligence*. Cambridge, 1991.