Abductive Inference with Probabilistic Graphical Models

Christian Borgelt

European Center for Soft Computing
c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Asturias, Spain
Email: christian.borgelt@softcomputing.es

Summary. Starting from a general characterization of logical inferences, I consider abductive reasoning, which aims at finding likely causes for observed symptoms. Such inferences are not truth preserving and thus it is necessary to assess their conclusions, to compare different explanations of the same findings, and finally to select the "best" hypothesis. Since in a large number of applications probabilistic graphical models are a mathematically sound and also very convenient tool for these operations, I discuss how they can be used to make abductive inference feasible.

1 Introduction

Abduction is a form of non-deductive logical inference, which aims at finding explanations for observations made (PEIR58; SALM73; CHAR97; PENG89; JOSE96; BORG00). On the other hand, probabilistic graphical models are a method to structure a multivariate probability distribution (mainly by finding a way to decompose it into distributions on lower-dimensional subspaces) and to compute efficiently (conditioned) marginal distributions on subspaces, especially individual variables (PEAR92; WHIT90; LAUR96; CAST97; JENS01; GAME04). Hence, at first glance, there seems to be little that these two areas have in common. However, the two notions are closely connected through (probabilistic) hypothesis assessment and statistical explanations.

In order to reveal this connection, I study a general model of abductive inference and hypothesis assessment. Unfortunately, this model is not suited for implementation, because it needs too much storage space and also because usually its parameters cannot be determined reliably. Direct approaches to simplify the model render it manageable, but require strong independence assumptions that are hardly acceptable in applications. Therefore a modeling technique is desired, by which we can take dependences between the involved variables into account, but which nevertheless lets us exploit (conditional) independences to simplify the model. One such technique, which has become very popular, are probabilistic graphical models. I review this modeling technique and discuss how it can be used for abductive inference.

2 Categorization of Logical Inferences

Logic, in the most general sense, describes the structure of languages in which one can argue. That is, logic is the (formal) theory of arguments, where an *argument* is a group of statements that are related to each other. More precisely, an argument consists of one statement representing the *conclusion* and one or more statements that give reasons supporting this conclusion. These latter statements are called *premisses* (SALM73). The process of deriving the conclusion from the premisses (using an argument) is called *inference.* Arguments are studied by analyzing the *inference rule* that is used. Such rules are usually stated in the form of *argument schemes.*

Łukasiewicz showed (according to (BOCH54)), that all logical inferences can be divided into two classes, which he called *deduction* and *reduction.* By exploiting logical equivalences we can modify the premisses of all arguments in such a way that arguments with only two premisses result. One of these premisses is a *conditional* or an *implication* (an if-then-statement), the other is equivalent either to the antecedent or to the consequent of this conditional[1]:

$$\text{Deduction:}\quad \begin{array}{l} A \rightarrow B \\ A \\ \hline B \end{array} \qquad\qquad \text{Reduction:}\quad \begin{array}{l} A \rightarrow B \\ B \\ \hline A \end{array}$$

Both of these inference rules are based on the tautology $((A \rightarrow B) \wedge A) \rightarrow B$, but they use it in different ways, which results in different properties.

Deduction serves the purpose to make all truths explicit that are determined by a set of statements. These truths are found by constructing appropriate arguments, which yield true conclusions, provided their premisses are true. Obviously, this holds only if no information is added to what is already present in the premisses. (If information was added, we could not guarantee the truth of the conclusion, because we would not know whether the additional information is correct.) Thus the basic properties of deduction are that it is infallible, but it does not tell us anything new (w.r.t. the premisses). These properties are a consequence of the fact that deductive inferences correspond exactly to tautologies (like, for example, the *modus ponens*).

Reduction serves the purpose to find explanations for statements that describe, for example, observations made. Obviously, the second premiss of a reductive argument can be obtained from the first premiss (the conditional) and the conclusion by a deductive inference. This is the rationale underlying them: the premiss B becomes a logical consequence and is thus “explained” by the conclusion. The drawback is that reduction is not truth preserving (which is not surprising, since information is added). Thus, the basic properties of reduction are that it is fallible, but as a compensation it tells us something new (w.r.t. the premisses). These properties are a consequence of the fact that there are no tautologies to which reductive inferences correspond directly.

[1] To avoid some technical problems, I implicitly assume throughout this paper that conditionals may or may not be (multiply) universally quantified.

Reduction can be further divided into induction and abduction.[2] This division is not based on the argument scheme, but on the types of statements involved: *Induction* goes from particular statements describing, for example, observations, symptoms, experiments etc., to general statements, that is, hypotheses and theories. *Abduction*, on the other hand, is a reductive inference in which the conclusion is a particular statement (BORG00). Using a notion of formal logic, we may also say that abduction is a reductive inference with the conclusion being a *ground formula*. That is, it must not contain variables, neither bound nor free, since they would render it a general statement.

Note that an analogous distinction is made for deduction, namely in the hypothetico-deductive method of science (HEMP66). There it is necessary, because only particular statements can be confronted with experimental findings; we can never observe a general law directly (POPP34). However, as far as I know, there are no special names for these two types of deductive inferences.

3 Hypothesis Assessment

A reductive—and thus an abductive—inference can yield a false conclusion, even if the premisses are true (due to the information added in the inference, see above). This is usually made explicit by calling the conclusion of a reductive argument a *hypothesis* or a *conjecture*. Therefore results of abductive inferences have to be assessed in order to minimize the chances that they are wrong.[3] The main criteria are (I do not claim that this list is complete):

- *Relation between antecedent and consequent*
  There must be a semantic connection between the antecendent and the consequent of the conditional, since otherwise the *material implication* used in formal logic allows for an abundance of meaningless explanations.
- *Relation to other statements*
  The inferred conclusion must be consistent with statements that are not used in the inference, but are known to be true (background knowledge, accepted theories). On the other hand, a hypothesis gets more plausible if several independent reductive arguments lead to it.
- *Parsimony (Ockham's razor)*
  An explanation should be as simple as possible, that is, should make as few and as simple assumptions as possible (*pluralitas non est ponenda sine necessitate*), since complex hypotheses are usually less likely to be true.

[2] There are other ways of defining abduction, which are discussed, for example, in (JOSE96; GABB00). However, I reject these alternative definitions due to reasons which are explained in detail in (BORG00).

[3] Note that most of the confusion about the meaning of the term "abduction" comes from a lack of distinction between the logical inference and the assessment of its result. Thus it is not surprising that there are as many different interpretations of the term "abduction" as there are criteria to assess hypotheses.

- *Probability*
  By exploiting prior and conditional probabilities we may sometimes be able to compute or at least estimate the probability of a hypothesis. This allows us to formally compare different hypotheses on solid grounds.

It should be noted that the last criterion (probability) captures a large part of the other criteria in a formal way. For example, parsimony can be seen as an intuitive and simplified probability assessment, since hypotheses combining several independent assumptions are usually less probable then simple ones making only a single assumption (as the probabilities of independent assumptions multiply). An exception are semantic considerations, which have to be treated, for instance, by fixing the set of useable conditionals and the set of *abducibles* (acceptable abductive hypotheses), maybe by a formal language.

4 Probabilistic Inferences

Up to now I assumed implicitly that the conditional (implication), which appears in deductive as well as in reductive arguments, is known to be absolutely correct (definite truth). However, in real world applications we rarely find ourselves in such a favorable position. To quote a well-known example: even the statement "If an animal is a bird, then it can fly." is not absolutely correct, since there are exceptions like penguins, ostriches etc. To deal with such cases—obviously, confining ourselves to absolutely correct conditionals would not be very helpful—we have to consider statistical syllogisms, statistical generalizations, and (particular) statistical explanations.

By *statistical syllogism* I mean a deductively shaped argument, in which the conditional is a statistical statement (like "80% of the beans in this box are white."). With a statistical conditional a deductively shaped argument loses its distinctive mark, that is, it is no longer infallible. Since the implication is not true in all cases (expressed by the associated percentage or probability), the conclusion may be false, even though the premisses are true. Nevertheless we can be rather confident that the conclusion of the argument is true if the conditional probability associated with the implication is fairly high. We may express this confidence by assigning to the conclusion a *degree of belief* (or a *degree of confidence*) equal to the probability stated in the conditional.

A *statistical generalization* is an argument that extends a statement about a sample to a statement about the whole population (like, for example, inferring the percentage of girls among newborn children from the observed percentage in a specific hospital). As a consequence, a statistical generalization is the probabilistic analog of an inductive argument. Statistical generalizations are one of the main topics of (inductive) statistics, where they are used, for instance, to predict the outcome of an election from polls. In (inductive) statistics it is studied what it takes to make the inferred statement reliable (theory of hypothesis testing), what the best estimates of parameters for the whole population are, and how to compute these best estimates.

I use the term *(particular) statistical explanation*[4] as a name for the probabilistic analog of an abductive inference. An example is a physician who infers the disease of a patient from observed symptoms and statistical knowledge about how often these symptoms are caused by different diseases the patient may have contracted. In order to assign a degree of confidence or degree of belief to the conclusion, we can draw on *Bayes' rule* in order to "invert" the conditional probability. That is, we compute

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where $P(B|A)$ is the conditional probability associated with the implication of the argument and $P(A)$ and $P(B)$ are the prior probabilities of the events in its antecedent and its consequent. Obviously, with Bayes' rule we can always translate a statistical explanation into a statistical syllogism, simply by turning the conditional around and associating it with the "inverted" conditional probability. Thus in probabilistic reasoning the difference between abductively and deductively shaped arguments vanishes—which is not surprising, since the distinctive mark of deductive inferences, their infallibility, is lost.

However, the other distinction made above, namely the distinction between arguments, the conclusion of which is a general statement, and arguments, the conclusion of which is a particular statement, remains valid. The reason is that it is next to impossible to know or even define the prior or posterior probability of a general statement (as is convincingly argued in (POPP34)). Therefore I confine myself to inferences of particular statements in the following, which is no real restriction, since my topic is abductive inference anyway.

5 General Model of Abductive Inference

In this section I briefly review a general model of abductive inference as it was presented in (BORG00). It incorporates a way of assessing hypotheses that is based on (BYLA91), but is also closely related to (PENG89). Since this model cannot be implemented directly (it would require too much storage space and usually its parameters cannot all be determined reliably) I also look for simplifications, which finally lead us to probabilistic graphical models.

5.1 Formal Definition

The following definition is intended to describe the framework in which the abductive reasoning takes place by fixing the statements that may be used in abductive arguments and a mechanisms to assess the resulting hypotheses.

[4] The word *particular* is added only to emphasize that the explanation must be a particular statement, since statistical generalizations—due to their reductive structure—also yield statistical explanations. However, statistical generalizations yield general statements, which refer to a whole population.

Definition 1. *An* **abductive problem** *is a tuple* $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$*, where*

- D_{all} *is a finite set of possible atomic data,*
- H_{all} *is a finite set of possible atomic hypotheses,*
- e *is a relation of* $2^{D_{\text{all}}}$ *and* $2^{H_{\text{all}}}$*, i.e.,* $e \subseteq 2^{D_{\text{all}}} \times 2^{H_{\text{all}}}$*,*
- pl *is a mapping from* $2^{D_{\text{all}}} \times 2^{H_{\text{all}}}$ *to a partially ordered set* Q*, and*
- $D_{\text{obs}} \subseteq D_{\text{all}}$ *is the set of observed data.*

The sets D_{all} and H_{all} contain the statements that can be used in inferences, the former the possible observations (or *data*), the latter the possible *hypotheses*. All statements in D_{all} and H_{all} are required to be particular, i.e., ground formulae. The two sets need not be disjoint, although this is useful for most applications. The relation e (for *explanation*) connects sets of observations with sets of hypotheses which explain them and thus describes the set of conditionals. The mapping pl (for *plausibility*) assesses the *quality* of an inferred (compound) hypothesis. In the following the set Q is always the interval $[0, 1]$, as I consider only probabilities and degrees of belief. D_{obs} is the set of *observed* data, that is, the set for which an explanation is desired.

Note that we may add to D_{all} statements that do not need an explanation, but may have a bearing on the assessment of possible hypotheses. For example, in medical diagnosis we register the sex of a patient, because the likelihood of certain diseases differs considerably for the two sexes. That is, D_{all} may contain not only general background knowledge about the application domain, but also relevant case-specific information.

An abductive problem is solved by finding the best explanation(s) for the data observed. This motivates the next two definitions.

Definition 2. *In an abductive problem* $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ *a set* $H \subseteq H_{\text{all}}$ *is called an* **explanation** *(of the data* D_{obs}*), iff* $(H, D_{\text{obs}}) \in e$*.*

Often an explanation is required to be parsimonious (see above). We may add this requirement by defining

$$H \text{ is an explanation, iff } \ (H, D_{\text{obs}}) \in e \ \wedge \ \neg\exists H' \subset H : (H', D_{\text{obs}}) \in e,$$

that is, H is an explanation only, if no proper subset of H is also an explanation. That is, we may require H to be *irredundant* (PENG89).

Definition 3. *In an abductive problem* $\mathcal{AP} = \langle D_{\text{all}}, H_{\text{all}}, e, pl, D_{\text{obs}} \rangle$ *an explanation* H *is called a* **best explanation** *(of the data* D_{obs}*), iff there is no explanation* H' *that is better than* H *w.r.t. the mapping* pl*, i.e., iff*

$$\neg\exists H' : (H', D_{\text{obs}}) \in e \ \wedge \ pl(H, D_{\text{obs}}) < pl(H', D_{\text{obs}}).$$

Of course, which explanation is selected depends on how the hypothesis assessment function pl ranks the possible explanations. Therefore it should be chosen with care. Fortunately, based on statistical arguments, we can identify the ideal choice, which may serve as a guideline.

Definition 4. *The* **optimal hypothesis assessment function** *is*

$$pl_{\text{opt}}(H, D) = P(H|D).$$

The best explanation under pl_{opt} is called the **most probable explanation***.*

The probability of the hypothesis given the data is the optimal hypothesis assessment function, because it is easy to show that choosing the hypothesis it advocates is, in the long run, superior to any other decision strategy—at least w.r.t. the relative number of times the correct decision is made. If the alternatives carry different costs in case of a wrong decision, an different function may be better. Nevertheless, the probability of the hypothesis given the data is still very important in this case, because it is needed to compute the hypothesis assessment function that optimizes the expected benefit.

The relation e and the mapping pl of an abductive problem can easily be represented as a table in which the rows correspond to the possible sets of hypotheses and the columns to the possible sets of observations (or vice versa). Each entry states the value assigned by the mapping pl to the pair (H, D) corresponding to the table field, provided that this pair is contained in the relation e (otherwise it is left null). With this representation solving an abductive problem is very simple: visit the table column that corresponds to the observed data and to find the row of this column that contains the highest probability. However, it is clear that for any real world problem worth considering we cannot set up this table, since it would have too many rows and columns. Therefore we have to look for simplifications, which exploit the structure of the relationships between the data and the possible hypotheses.

5.2 Simplifications

Let us consider, in two steps, simplifications of the general model. The first is based on the idea to replace the relation e by a function, which assigns to a set H of hypotheses the union of all sets D of observations that H can explain. Of course, this requires specific conditions to hold.

Definition 5. *An abductive problem is called* **functional** *iff*

1. $\forall H \subseteq H_{\text{all}}: \quad \forall D_1, D_2 \subseteq D_{\text{all}}:$
   $((H, D_1) \in e \wedge D_2 \subseteq D_1) \Rightarrow (H, D_2) \in e$
2. $\forall H \subseteq H_{\text{all}}: \quad \forall D_1, D_2 \subseteq D_{\text{all}}:$
   $((H, D_1) \in e \wedge (H, D_2) \in e \wedge D_1 \cup D_2 \textit{ is consistent}) \Rightarrow (H, D_1 \cup D_2) \in e$

Note that the first condition is no real restriction, since sets of observations are interpreted as conjunctions. The second condition is a restriction, though, since counterexamples can easily be found (see, for instance, (BORG00)). But if these conditions hold, the relation e can be replaced by a function e_f, where

$$\forall H \subseteq H_{\text{all}}: \quad e_f(H) = \{d \in D_{\text{all}} \mid \exists D \subseteq D_{\text{all}} : d \in D \wedge (H, D) \in e\}.$$

With this function e_f an explanation is defined as a set $H \subseteq H_{\mathrm{all}}$, such that $D_{\mathrm{obs}} \subseteq e_f(H)$. Hence it allows to represent the explanation relation by a table with one column for each $d \in D_{\mathrm{all}}$ and one row for each $H \subseteq H_{\mathrm{all}}$. However, in order to represent the plausibility assessment function *pl* with a similar table (otherwise we do not gain anything), we need even stronger assumptions.

Definition 6. *A functional abductive problem is called* **D-independent**, *iff*

$$\forall H \subseteq H_{\mathrm{all}} : \forall D \subseteq D_{\mathrm{all}} : \quad D \textit{ is consistent} \quad \Rightarrow \quad P(D|H) = \prod_{d \in D} P(d|H).$$

Intuitively, D-independence means that the probability of a possible observation d is independent of any other observations that may be present given any set of hypotheses H. Under these conditions we can compute the assessment of any consistent set H of hypotheses (using Bayes' rule) as

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = P(H) \prod_{d \in D} \frac{P(d|H)}{P(d)},$$

provided we also know the prior probability $P(H)$. This approach, although restricted to one element sets $\{h\}$, was suggested in (CHAR97). For general sets H, consisting of several individual hypotheses, it is still not usable, since the needed table has too many rows. Therefore we need further simplifications.

Definition 7. *A functional abductive problem is called* **independent**, *iff*

$$\forall H \subseteq H_{\mathrm{all}} : \quad e_f(H) = \bigcup_{h \in H} e_f(\{h\}),$$

where e_f is the function by which the relation e can be represented in a functional abductive problem (see above).

Of course, this is a very strong restriction. It requires that there is no interaction of hypotheses in the sense that no combination of atomic hypotheses can explain an observation, which cannot be explained by at least one of the contained atomic hypothesis alone. In addition, the explanatory powers of atomic hypothesis must not cancel each other (there must not be "destructive interference"), so that a combination of hypotheses can no longer explain what one of them could explain individually. Unfortunately, this requirement excludes fairly commonplace situations (see (BORG00) for examples). Even worse, this assumption alone is not even enough to allow us to simplify the plausibility assessment function *pl* in a similar way. In order to achieve a simplification of the function *pl*, we need even stronger assumptions.

Definition 8. *A D-independent abductive problem is* **HD-independent**, *iff*

$$\begin{array}{l} \forall H \subseteq H_{\mathrm{all}} : \forall d \in D_{\mathrm{all}} : \\ H \textit{ is consistent} \ \Rightarrow \ P(H|d) = \prod_{h \in H} P(h|d) \ \wedge \ P(H) = \prod_{h \in H} P(h). \end{array}$$

As a result of this final assumption, we only need to store the probabilities $P(h)$, $P(d)$, and $P(d|h)$ for all $h \in H_{\rm all}$ and all $d \in D_{\rm all}$, that is, only $(|H_{\rm all}|+1)\cdot(|D_{\rm all}|+1)-1$ probabilities. The conditional probability of a set H of hypotheses given a set D of possible observations is computed as:

$$P(H|D) = \prod_{h' \in H} P(h') \prod_{d \in D} \left(\prod_{h \in H} \frac{P(d|h)}{P(d)} \right).$$

This model, which was derived in (BORG00), is surely feasible in terms of storage and computation requirements. However, it is unlikely to be useful in practice, due to its fairly extreme independence assumptions. However, studying the simplifications that led to it reveals fairly clearly where the problems of abductive inference lie. Without exploiting (conditional) independences, we cannot reach a feasible model. However, assuming too much independence renders the model too strict and thus inappropriate for practical purposes.

6 Probabilistic Graphical Models

In order to cope with the problem of achieving simplifications without introducing extreme independence assumptions, we may search for a model, in which we can take dependences into account, but nevertheless can exploit all existing independences to reduce the amount of storage needed and to make inferences tractable. Probabilistic graphical models are such an approach.

6.1 Decomposition and Abductive Reasoning

Concisely stated, the basic ideas underlying probabilistic graphical models are these: under certain conditions a probability distribution P on a multi-dimensional domain, which encodes *prior knowledge* about this domain, can be decomposed into a set $\{P_1, \ldots, P_n\}$ of probability distributions on lower-dimensional subspaces. This decomposition is based on dependence and independence relations between the attributes used to describe the domain. If a decomposition is possible, it is sufficient to know the distributions on the subspaces to compute all probabilities that can be computed using the original distribution P. Since such a decomposition can be represented as a network (or graph), it is commonly called a *probabilistic network* or a *probabilistic graphical model.* Reasoning in such a network consists in conditioning the represented probability distribution w.r.t. the observed values of some attributes.

A decomposition of a probability distribution has several advantages. The most important are that it can usually be stored much more efficiently and with less redundancy than the original distribution. However, this would be of little use for reasoning tasks, were it not for the possibility to draw inferences using only the distributions $\{P_1, \ldots, P_r\}$ on the subspaces without having to reconstruct the original distribution P. If we obtained *evidential knowledge* about the current state of the domain under consideration, which consists of observed values for some of the attributes, we can *condition* the represented probability distribution on the observed values by passing the conditioning information from subspace distribution to subspace distribution until all have been updated. This process is usually called *evidence propagation.*

Mapping the abductive inference model to probabilistic networks is—for the greater part—very simple. In the first place, we form groups of mutually exclusive and exhaustive statements from the statements in H_{all} and D_{all}, which then form the dimensions of the multidimensional space. The hypothesis assessment function *pl* corresponds directly to the probability distribution P on the domain constructed in this way, since from this probability distribution we can compute the probability $P(H|D)$ for all sets H and D. Hence the decomposition can be used to simplify the representation of the hypothesis assessment function *pl*. The observed data D_{obs} corresponds, of course, to the evidential knowledge. The only element of an abductive problem for which there is no direct analog is the explanation relation e (see below).

6.2 Conditional Independence

Whether and how a given probability distribution P can be decomposed into a set $\{P_1, \dots, P_r\}$ of distributions on subspaces is determined by the dependence structure of the attributes of the domain Ω underlying P. The core notion governing this decomposition is that of attributes being *(probabilistically) conditionally independent* of each other (DAWI79; PEAR92).

Definition 9. *Let P be a probability distribution on the space spanned by the attributes in $V = \{A_1, \dots, A_m\}$ and let X, Y, and Z be three disjoint subsets of attributes in V. X is called* **conditionally independent** *of Y given Z w.r.t. P, written $X \perp\!\!\!\perp Y \mid Z$, iff whenever $P(\omega_Z) > 0$ we have*

$$\forall \omega \in \Omega : \quad P(\omega_{X \cup Y} \mid \omega_Z) = P(\omega_X \mid \omega_Z) \cdot P(\omega_Y \mid \omega_Z),$$

where ω_X, ω_Y etc. are instantiation of the variables in X, Y etc., respectively.

It has been shown in general that a notion of conditional independence satisfying certain axioms, which are known as the *semi-graphoid axioms* (DAWI79; PEAR92), can be used to define a graph structure on the set of attributes. These axioms are:

symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$

decomposition: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z)$

weak union: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z \cup W)$

contraction: $(W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

The *symmetry* axiom states that in any state of knowledge Z, if Y tells us nothing new about X, then X tells us nothing new about Y. The *decomposition* axiom asserts that if two combined items of information are judged irrelevant to X, then each separate item is irrelevant as well. The *weak union* axiom states that learning irrelevant information W cannot help the irrelevant information Y become relevant to X. The *contraction* axiom says that if we judge W irrelevant to X after learning irrelevant information Y, then W must have been irrelevant before. Together the weak union and contraction properties mean that irrelevant information should not alter the relevance of other propositions in the system; what was relevant remains relevant, and what was irrelevant remains irrelevant (PEAR92). It is plausible that a reasonable notion of conditional independence should satisfy these axioms and, indeed, probabilistic conditional independence does.

6.3 Graph Representation

The notion of conditional independence provides the connection to a graph representation. In a *conditional independence graph* $G = (V, E)$ for a given probability distribution P each node represents an attribute of the underlying domain. The topology of the graph (i.e., which edges are present and which are missing) is an independence model of the distribution P. In particular, the graph represents a set of conditional independence statements about the distribution P by a notion of *node separation* (PEAR92; SPIR93).

What is to be understood by "separation" depends on whether the graph is directed or undirected. If X, Y, and Z are three disjoint subsets of nodes in an undirected graph, then Z separates X from Y, iff after removing the nodes in Z and their associated edges from the graph there is no path, i.e., no sequence of consecutive edges, from a node in X to a node in Y. Or, in other words, Z separates X from Y, iff all paths from a node in X to a node in Y contain a node in Z. For directed graphs, which have to be acyclic, the so-called *d-separation criterion* is used (PEAR92; VERM90): If X, Y, and Z are three disjoint subsets of nodes in a directed acyclic graph (DAG), then Z is said to *d-separate* X from Y, iff there is no path, i.e., no sequence of consecutive edges (of any directionality), from a node in X to a node in Y along which the following two conditions hold:

1. every node with converging edges either is in Z or has a descendant in Z,
2. every other node is not in Z.

With the described two notions of node separation, we can define the so-called *Markov properties* of graphs (WHIT90). For example, for undirected graphs, these properties are defined as follows (for directed graphs they are similar):

pairwise: Attributes, whose nodes are non-adjacent in the graph, are conditionally independent given all remaining attributes.

local: Given the attributes of the adjacent nodes (the neighbors), an attribute is conditionally independent of all remaining attributes.

global: Any two subsets of attributes, whose corresponding node sets are separated by a third node set, are conditionally independent given the attributes corresponding to the nodes in the third set.

Note that the local Markov property is contained in the global, and the pairwise Markov property in the local. However, the three types are not equivalent in general, and it is obvious that we need the *global* Markov property for inferences from multiple observations. However, the above definition can be used if—in addition to the semi-graphoid axioms—the following axiom holds:

intersection: $(W \perp\!\!\!\perp Y \mid Z \cup X) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Longrightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

The semi-graphoid axioms together with this one are called the *graphoid axioms.* If can be shown that a strictly positive probability distribution satisfies the intersection axiom (PEAR92) and therefore the probability distribution on the modeled domain is often required (or assumed) to be strictly positive.

6.4 Factorization

The conditional independence graph is also called the *qualitative* part of a probabilistic graphical model, since it specifies which attributes are dependent and which are (conditionally) independent, but not the exact details of the dependences. These details are represented in the *quantitative* part of a probabilistic graphical model. It consists of a set of probability distributions and describes a *factorization* of the joint probability distribution P. The exact representation of the quantitative information and the factorization formula depends on the type of the conditional independence graph.

- *Bayesian networks*
  The most popular probabilistic network is the *Bayesian network*, also called *belief network* or (somewhat misleadingly) *probabilistic causal network.* It consists of a directed acyclic graph and a set of conditional probability distributions $P(\omega_A \mid \omega_{\text{parents}(A)})$, $A \in V$, where parents(A) is the set of attributes corresponding to the parents of the node that corresponds to attribute A. A Bayesian network describes the factorization

$$\forall \omega \in \Omega: \quad P(\omega) = \prod_{A \in V} P(\omega_A \mid \omega_{\text{parents}(A)}).$$

- *Markov networks*
  An alternative type of probabilistic networks uses undirected graphs and is called a *Markov network.* It represents so-called *Markov random fields.* Similar to a Bayesian network a Markov network describes a factorization of the joint probability distribution P, but it uses a *potential representation*: a strictly positive probability distribution P factorizes w.r.t. an undirected graph $G = (V, E)$, iff

$$\forall X \in \text{cliques}(G) : \exists \phi_X : \forall \omega \in \Omega : \quad P(\omega) = \prod_{X \in \text{cliques}(G)} \phi_X(\omega_X),$$

  where cliques(G) is the set of all maximal cliques of G. The *factor potentials* ϕ_X are strictly positive functions defined on Ω_X, $X \subseteq V$, which can be computed from the corresponding marginal distributions.

6.5 Evidence Propagation

After a probabilistic network has been constructed, it can be used to do reasoning. Although this is fairly straightforward in general, considerations of efficiency make it often advisable to transform a graphical model into a form that is better suited for propagating the evidential knowledge and computing the resulting marginal distributions for the unobserved attributes. Among the most popular techniques is *clique tree propagation* (CTP) (LAUR88; CAST97; JENS01), which involves transforming the conditional independence graph into a clique tree. Alternatives include bucket elimination (DECH96; ZHAN96) and iterative proportional fitting (WHIT90). Commonly used evidence propagation algorithms differ from each other w.r.t. the network structures they support. For example, bucket elimination can also be used with networks that contain cycles, while other methods need cycles to be eliminated, temporarily "cut open" or treated with other special techniques.

6.6 Hypothesis Selection

As indicated above, probabilistic networks provide means to represent the hypothesis assessment function pl for abductive reasoning. However, in a probabilistic network there is no direct analog to the explanation relation e, which identifies the relevant hypotheses. That is, with a probabilistic network we can compute the plausibility of a given hypothesis, but we cannot check whether the hypothesis is semantically acceptable (which is the main purpose of the explanation relation e). Fortunately, in models tailored for a specific application this is often irrelevant, because from the application it can be clear what attributes we are interested in and therefore we only have to compute the most probable tuple of values for the subspace formed by these attributes.

To identify the most probable tuple in a subspace formed by a set of attributes, may not always be appropriate, though. If, for instance, not all acceptable (compound) hypotheses consist of the same number of atomic hypotheses, we cannot use it, because it may result in hypotheses that are too specific for a given problem. However, even in this case the probabilistic network alone may contain enough information to select the best acceptable hypothesis. For example, the structure of the network can provide information how to restrict the set of attributes we have to take into account to form a (compound) hypothesis. Obviously, it is sufficient to select a set of explanatory attributes (i.e., attributes derived from H_{all}) in such a way that the observed attributes and the remaining attributes are conditionally independent given the selected attributes. A (compound) hypothesis formed from these attributes has to be considered complete, because due to the interpretation of the semi-graphoid axioms (see above) the remaining attributes are irrelevant for the observations. Of course, such a restriction does not guarantee that the selected (compound) hypothesis is semantically acceptable, but it may help to restrict the set of hypotheses one has to consider. The idea can be enhanced by the requirement that a reasonable hypothesis should make the observed data more likely than it is without it or that the observed data should make an acceptable hypothesis more likely (even though this is not sufficient to make a hypothesis semantically acceptable).

If these approaches, which try to do without additional information, are not feasible, we have to add some structure to represent (a simplification of) the explanation relation e. A very manageable structure results if we have an independent abductive problem and thus can represent the relation e as a simple table with H_{all} lines and D_{all} columns. In this case the probabilistic network helps to avoid the strong probabilistic independence assumptions underlying D-independent and HD-independent abductive problems. We only need the (weaker) logical independence assumptions needed to simplify the representation of the relation e. If even these logical assumptions are too strong, we can enhance the table of the relation e for an independent abductive problem by an explicit list of (compound) hypotheses, for which “constructive” or “destructive inference” occurs, i.e., those (compound) hypotheses which can explain more than the sum of their elements and those, which can explain less. Provided this list is of moderate size, the problem remains tractable.

6.7 Learning from Data

A probabilistic network is a powerful tool to support reasoning—as soon as it is constructed. Its construction by human experts, however, can be tedious and time consuming. Therefore a large part of research in probabilistic graphical models focused on learning them from a database of sample cases. In accordance with the two components of graphical models, one distinguishes between *quantitative* and *qualitative* (or *structural*) *network induction.*

- *Quantitative network induction*
  Given a graph, the parameters of the conditional probability distributions or the factor potentials are estimated. A lot of approaches have been developed in this field, using methods such as maximum likelihood, maximum penalized likelihood, or fully Bayesian approaches.
- *Qualitative network induction*
  The graph underlying the network is induced from a database of sample cases. The most popular approaches are based on conditional independence tests (CI tests) (VERM92; CHEN02) and on Bayesian inference (COOP92; HECK95). All of them work reasonably well in practice, but still suffer from some problems.

More details about learning probabilistic graphical models from data can be found in (CAST97; JORD98; BORG02). The last of these references also considers learning possibilistic graphical models, which use an alternative uncertainty calculus and can be useful in scenarios where the available information is insufficient or does not allow for a proper estimation of probabilities.

7 Summary

In this paper I considered how probabilistic networks can support abductive reasoning. Starting from a definition of an abductive inference as a reductive, i.e., explanatory inference, the conclusion of which is a particular statement, I showed how probability theory enters the consideration due to two reasons: in the first place, if we want to handle real world problems, we have to take into account statistical conditionals. Secondly, in order to reduce the chances of an incorrect result, we have to assess and compare the conclusions of abductive inferences. Based on a general model of abductive inference I showed that a direct approach to represent a hypothesis assessment function is not feasible and thus simplifications are required. Although straightforward simplifications lead to a manageable model, they involve strong assumptions which cannot reasonably be expected to hold in applications. As a solution probabilistic networks suggest themselves as a well-established technique to decompose a multivariate probability distribution in order to make reasoning in high-dimensional domains possible. They are very well-suited to represent the hypothesis assessment function of abductive problem solving. However, it may be necessary to enhance them by a method to identify the (semantically) acceptable hypotheses, because the raw probabilistic information they represent is often not sufficient for this task.

References

[BOCH54] Bocheński, I.M.: *Die zeitgenössischen Denkmethoden.* Franke-Verlag, Tübingen, Germany 1954

[BORG00] Borgelt, C. and Kruse, R.: Abductive Inference with Probabilistic Networks. In (GABB00), 281–314.

[BORG02] Borgelt, C. and Kruse, R.: *Graphical Models — Methods for Data Analysis and Mining.* J. Wiley & Sons, Chichester, United Kingdom 2002

[BYLA91] Bylander, T., Allemang, T., Tanner, M.C. and Josephson, J.R.: The Computational Complexity of Abduction. *Artificial Intelligence* 49: 25–60. North-Holland, Amsterdam, Netherlands 1991

[CAST97] Castillo, E., Gutierrez, J.M. and Hadi, A.S.: *Expert Systems and Probabilistic Network Models.* Springer, New York, NY, USA 1997

[CHAR97] Charniak, E. and McDermott, D.: *Introduction to Artificial Intelligence.* Addison-Wesley, Reading, MA, USA 1985

[CHEN02] Cheng, D., Greiner, D., Kelly, J., Bell, D.A. and Liu, W.: Learning Bayesian Networks from Data: An Information Theory Based Approach. *Artificial Intelligence* 137(1–2):43–90. Elsevier, Amsterdam, Netherlands 2002

[COOP92] Cooper, G.F. and Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning* 9:309–347. Kluwer, Dordrecht, Netherlands 1992

[DAWI79] Dawid, A.: Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society (Series B)* 41:1–31. Blackwell, Oxford, United Kingdom 1979

[DECH96] Dechter, R.: Bucket Elimination: A Unifying Framework for Probabilistic Inference. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI'96, Portland, OR)*, 211–219. Morgan Kaufman, San Mateo, CA, USA 1996

[GABB00] Gabbay, D.M. and Kruse, R. eds.: *Abductive Reasoning and Learning*, volume 4 of *Handbook of Defeasible Reasoning and Uncertainty Management Systems.* Kluwer, Dordrecht, Netherlands 2000

[GAME04] Gamez, J.A., Moral, S. and Salmeron, A.: *Advances in Bayesian Networks.* Springer, Berlin, Germany 2004

[HECK95] Heckerman, D., Geiger, D. and Chickering., D.M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243. Kluwer, Dordrecht, Netherlands 1995

[HEMP66] Hempel, C.G.: *Philosophy of Natural Science.* Prentice-Hall, Englewood Cliffs, NJ, USA 1966

[JENS01] Jensen, F.V. *Bayesian Networks and Decision Graphs.* Springer, Berlin, Germany 2001

[JORD98] Jordan, M.I. ed.: *Learning in Graphical Models.* MIT Press, Cambridge, MA, USA 1998

[JOSE96] Josephson, J.R. and Josephson, S.G.: *Abductive Inference — Computation, Philosophy, Technology.* Cambridge University Press, Cambridge, MA, USA 1996

[LAUR88] Lauritzen, S.L. and Spiegelhalter, D.J.: Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, 2(50):157–224. Blackwell, Oxford, United Kingdom 1988

[LAUR96] Lauritzen, S.L.: *Graphical Models.* Oxford University Press, Oxford, United Kingdom 1996

[PEAR92] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (2nd edition).* Morgan Kaufman, San Mateo, CA, USA 1992

[PEIR58] Peirce, C. S. (Hartshorne, C., Weiss, P. and Burks, A., eds.): *Collected Papers of Charles Sanders Peirce.* Havard University Press, Cambridge, MA, USA 1958

[PENG89] Peng, Y. and Reggia, Y. *Abductive Inference Models for Diagnostic Problem Solving.* Springer, New York, NY, USA 1989

[POPP34] Popper, K.R.: *Logik der Forschung.* 1st edition: J. Springer, Vienna, 1934. 9th edition: J.C.B. Mohr, Tübingen, 1989. English edition: *The Logic of Scientific Discovery*, Hutchinson, London, United Kingdom 1959

[SALM73] Salmon, W.C.: *Logic (2nd edition).* Prentice-Hall, Englewood Cliffs, NJ, USA 1973

[SAVA54] Savage, L.J.: *The Foundations of Statistics.* J. Wiley & Sons, New York, NY, USA 1954

[SPIR93] Spirtes, P., Glymour, C. and Scheines, R.: *Causation, Prediction, and Search (Lecture Notes in Statistics 81).* Springer, New York, NY, USA 1993

[VERM90] Verma, T.S. and Pearl, J.: Causal Networks: Semantics and Expressiveness. In: R.D. Shachter, T.S. Levitt L.N. Kanal, and J.F. Lemmer, eds. *Uncertainty in Artificial Intelligence 4*, pp. 69–76. North Holland, Amsterdam, Netherlands 1990

[VERM92] Verma, T.S. and Pearl, J.: An Algorithm for Deciding if a Set of Observed Independencies has a Causal Explanation. *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence (UAI'92, Stanford, CA)*, pp. 323–330. Morgan Kaufman, San Mateo, CA, USA 1992

[WHIT90] Whittaker, J.: *Graphical Models in Applied Multivariate Statistics.* J. Wiley & Sons, Chichester, United Kingdom 1990

[ZHAN96] Zhang, N. L. and Poole, D.: Exploiting Causal Independence in Bayesian Network Inference. *Journal of Artificial Intelligence Research* 5:301–328. Morgan Kaufman, San Mateo, CA, USA 1996