

Fundamentals of Canonical Models

F. J. Díez
Dpto. Inteligencia Artificial
UNED
Senda del Rey, 9
28040 Madrid, Spain
fjdiez@dia.uned.es

Marek J. Druzdzel
Decision Systems Laboratory
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260, U.S.A.
marek@sis.pitt.edu

Abstract: *Canonical models are useful not only because they simplify the construction of probabilistic models, but also because they save storage space and computational time, and because they respond to causal patterns that can be exploited to generate user explanations. In this paper we offer a general framework for canonical models and briefly analyze the properties of the OR/MAX family of models. The general framework can be easily used to generate other canonical models.*

Keywords: Bayesian networks, canonical probabilistic models, knowledge engineering

1 Introduction

The construction of graphical probabilistic models, such as Bayesian networks and influence diagrams, requires the specification of many conditional probability distributions of the form $P(y|\mathbf{x})$, where $\mathbf{X} = \{X_1, \dots, X_n\}$ is the set of parents of node Y in the network. Today's graphical models almost always use only discrete variables, due to the difficulty of representing continuous probability distributions (except for a few cases) and the difficulty of propagating evidence. Therefore, probability distributions are usually given in the form of conditional probability tables (CPT's). In general, the numerical parameters are obtained from databases or assessed by human experts, and for this reason it is usually difficult to build a CPT for a family having more than two or three parents. In case of a database, the difficulty arises when a certain configuration is not represented in the database. For instance, when \mathbf{X} represents the set of diseases that may cause a certain anomaly Y , and \mathbf{x} is a particular configuration corresponding to the presence of several infrequent diseases, it is quite likely that the database contains no patient for that configuration. In case a knowledge engineer asks a human expert to estimate the conditional

probabilities, the difficulty of estimating the probability of infrequent configurations is even more serious, and there also additional difficulties when the number of probabilities to be estimated is high, because in general the time available for interaction with the expert is scarce. For these reasons it is desirable to dispose of *canonical models* that allow to build large CPT's from a small number of parameters that correspond to frequent configurations. The term "canonical" is used because such models are elementary units used in the construction of more complicated models [7].

Canonical models are useful not only because they simplify the construction of probabilistic models (knowledge engineering), but also because they save storage space and computational time [2] and because they respond to causal patterns that can be exploited to generate user explanations [5]. Although canonical models are more and more used in probabilistic expert systems, we believe that their advantages have not been fully explored yet, and for this reason this paper tries to offer two contributions to the study of canonical models: *(i)* a general framework, structured in three categories: deterministic, noisy, and leaky models, which provides a unified description of existing models and a tool for building new ones when they are required in specific domains, and *(ii)* an analysis of the properties of the models in the OR/MAX family, with a special emphasis on knowledge engineering issues. Sections 2 and 3 are devoted to these two topics, respectively, and Section 4 to the conclusions.

1.1 Definitions and notation

We will use capital letters to represent variables and lowercase letters to represent their values; for instance, v will represent any of the values of variable V . In the same way, \mathbf{V} will denote a set of variables $\{V_1, \dots, V_n\}$ and \mathbf{v} a certain n -tuple (v_1, \dots, v_n) , where v_i represents the value taken on by variable V_i .

There are some binary variables such that one outcome represents the presence of something (in diagnostic problems, the presence of an anomaly or disease) or a positive result in a test, and the other outcome represents the absence of the same entity or a negative result in the test. In this case, we denote the first value by 1 or $+v$ and the other by 0 or $\neg v$ and establish that $+v > \neg v$.

Given a set \mathbf{V} of binary variables of type present/absent or positive/negative, for each configuration \mathbf{v} we define \mathbf{V}^+ as the subset of variables taking on value 1 in such configuration (in diagnosis problems, we can view \mathbf{V}^+ as the set of anomalies present in the particular case we are considering):

$$\mathbf{V}^+ = \{V_i \in \mathbf{V} | v_i = 1\} \quad (1)$$

We also define $+\mathbf{v}$ as the configuration in which all variables take the value 1:

$$+\mathbf{v} = (+v_1, \dots, +v_n) \quad (2)$$

and $\neg\mathbf{v}$ as the configuration in which all variables take the value 0:

$$\neg\mathbf{v} = (\neg v_1, \dots, \neg v_n) \quad (3)$$

2 General properties of canonical models

Every canonical model represents a probabilistic relation involving a finite number of nodes, such that Y is the *child* and $\{X_1, \dots, X_n\}$ are its *parents* (see Figure 1). This terminology proceeds from assuming that the family makes part of a Bayesian network, but these models can also be imbedded in other probabilistic formalisms, such as Markov networks and chain graphs. Of course, different canonical models may coexist in a certain probabilistic model; in real-world applications it is typical to have Bayesian networks in which some of the families interact through OR-gates, a few through AND-gates and the rest of the families do not correspond to any canonical model, which implies that their CPT must be explicitly given. All the variables involved in the models that we consider in this paper are discrete variables; we also assume that the number of values of each variable is finite.

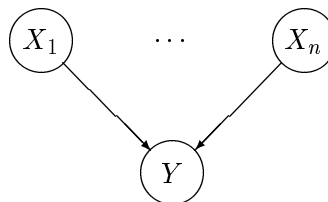


Figure 1. Node Y has n parents, X_1, \dots, X_n .

In the present paper we describe these models as different types of **causal interactions**. The reason is that each model tries to mimic a certain pattern of behavior that we observe or assume between causes and effects in the real world; this facilitates our task of explaining each model, defining its parameters and offering criteria for applying them. Nevertheless, it is also possible to view these models as mere probabilistic relations, without any causal interpretation.

2.1 Deterministic models

Some of these models are deterministic; in this case, the value taken on by Y is a function of the values of the X_i 's: $y = f(x_1, \dots, x_n)$, and the CPT is given by

$$P(y|\mathbf{x}) = \begin{cases} 1 & \text{if } y = f(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Please note that the deterministic models do not need any parameter: the CPT can be obtained just from f by the above equation.

Therefore, a deterministic model is characterized by

- **The domains of the variables, Y and X_i 's:** for instance, all the nodes involved in the OR, AND, and XOR models correspond to binary variables of type

present/absent or positive/negative; in contrast, the MAX and MIN models only require that Y is an ordinal variable.

- **The function f :** in some of these models, f corresponds to a logical operator, such as NEG, OR, AND or XOR, from which the model takes its name. In electronic circuits, these operators are implemented by elementary devices called *gates*; for analogy, some of this probabilistic models are often termed OR-gate, AND-gate, MAX-gate, etc.

The functions we discuss in this paper accept a finite but indefinite number of arguments and are all commutative.

2.2 Noisy models

We also analyze in this paper noisy models, which are built by introducing n auxiliary variables $\{Z_1, \dots, Z_n\}$ (Figure 2), such that the interaction between Y and the Z_i 's corresponds to a deterministic model (see the above section) and each relation between X_i and Z_i is given by a CPT $P(z_i|x_i)$. These auxiliary Z_i 's are used for deriving the equations, but do not make part of the final model, whose CPT is obtained by marginalizing out the Z_i 's:

$$P(y|\mathbf{x}) = \sum_{\mathbf{z}} P(y|\mathbf{z}, \mathbf{x}) \cdot P(\mathbf{z}|\mathbf{x}) \tag{5}$$

Because of Equation 4 and the Markov property,

$$P(y|\mathbf{x}) = \sum_{\mathbf{z}|f(\mathbf{z})=y} \prod_i P(z_i|x_i) \tag{6}$$

This equation is valid for all noisy models, although there are more efficient ways of computing the CPT for some models, such as the OR, AND, MAX, and MIN.

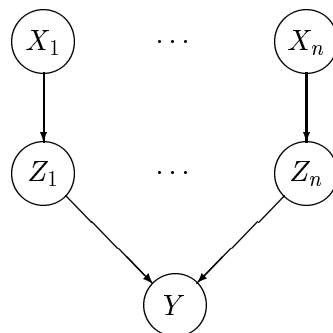


Figure 2. Auxiliary structure for the derivation of a noisy model.

In some models, the Z_i 's have a causal interpretation. For instance, in the noisy OR Z_i represents whether effect Y has been produced by cause X_i .

Some models include restrictions on the values of $P(z_i|x_i)$. For instance, in the noisy OR, $P(+z_i|\neg x_i) = 0$, which means that cause X_i , when absent, cannot produce effect Y . Similar restrictions apply to the AND, MAX and MIN models.

A specific feature of noisy models is that each parameter $P(z_i|x_i)$ is associated with a particular link $X_i \rightarrow Y$, while every parameter $P(y|\mathbf{x})$ in a CPT corresponds to a certain configuration \mathbf{x} made up by all the parents of Y , and cannot be associated to any particular link. Two advantages of canonical models arise from these properties. The first one is the reduction in the number of parameters from $O(\exp(n))$ in the case of a CPT to $O(n)$ for a canonical model, which is a substantial improvement from the point of view of the knowledge elicitation. The second advantage is that the parameters in canonical models lend themselves to more intuitive interpretations, which facilitates the task of estimating them.

2.3 Leaky models

In practical applications, it is often unfeasible to list all the possible parents of a certain node Y . In this case, we can assume that there is a large Bayesian network that properly represents the real-world domain. In this hypothetical network, the set of parents of Y is $\mathbf{X}^\dagger = \{X_1, \dots, X_{n^\dagger}\}$, and the interaction between \mathbf{X}^\dagger and Y is given by a certain canonical model. We may decide to build a simplified Bayesian network in which only a subset $\mathbf{X} = \{X_1, \dots, X_n\}$ of the parents are explicitly represented ($\mathbf{X} \subset \mathbf{X}^\dagger$); the rest of the parents, $\mathbf{X}_I = \mathbf{X}^\dagger \setminus \mathbf{X} = \{X_{n+1}, \dots, X_{n^\dagger}\}$, the ancestors of \mathbf{X}_I and the descendants of \mathbf{X}_I (except for Y and its descendants) are not explicitly represented in the simplified network.

In this case, the interaction between Y and \mathbf{X} in the simplified network can be represented by a leaky noisy model, provided that some conditions hold. The first condition is that the function f that defines the canonical model in the large network model must be associative, in the sense that when $n \geq 3$,

$$f(x_1, \dots, x_n) = f(x_1, f(x_2, \dots, f_n)) \tag{7}$$

There are other graphical conditions that we do not discuss in this paper, due to the lack of space.

The CPT for Y in the simplified network is

$$P(y|\mathbf{x}) = \sum_{\mathbf{z}} \left(\prod_i P(z_i|x_i) \right) \cdot \left(\sum_{y_L|f(\mathbf{z}, y_L)} P(y_L) \right) \tag{8}$$

where $P(y_L)$ is a (vectorial) parameter of the leaky model. In principle, this parameter could be computed from the large network, but in practice it is usually extracted from a database or estimated by an expert.

Any marginal or conditional probability obtained from the simplified network is the same as the one we would obtain from the large network.

3 Specific models

3.1 OR models

3.1.1 Deterministic OR

The deterministic OR model is a probabilistic relation among $n + 1$ binary variables, all taking values in $\{0, 1\}$. The function that defines it is

$$y = f_{\text{OR}}(\mathbf{x}) = \begin{cases} \neg y & \text{if } \mathbf{x} = \neg \mathbf{x} \\ +y & \text{otherwise} \end{cases} \quad (9)$$

This means that Y is absent only when all the X_i 's are absent, and Y is present when any of the X_i 's is present:

$$+y \longleftrightarrow (+x_1 \vee \dots \vee +x_n) \quad (10)$$

This is the reason why this model is termed OR gate.

3.1.2 Noisy OR

The noisy OR was proposed by Good [3] and further studied by Pearl [6, 7]. It can be obtained from the deterministic OR by introducing n auxiliary variables $\{Z_1, \dots, Z_n\}$, as discussed in Section 2.2 (see Fig. 2); the relation between Y and the Z_i 's is given by a deterministic OR, so that $y = f_{\text{OR}}(\mathbf{z})$.

The causal interpretation of this model is that each X_i represents one of the causes of Y . The term “noisy” accounts for the possibility that some of the causes fail to produce the effect even when they are present. Each Z_i indicates whether X_i has produced Y ; then, $\neg z_i$ means that X_i has not produced Y , either because X_i is absent or because a certain inhibitor has prevented X_i from producing Y . If we denote by q_i the probability that inhibitor I_i is active, then the probability c_i that X_i produces Y when it is present is:

$$c_i = P(+z_i | +x_i) = 1 - q_i \quad (11)$$

Naturally, when X_i is absent, it can not cause Y :

$$P(+z_i | \neg x_i) = 0 \quad (12)$$

The deterministic OR is a particular case of the noisy OR in which $c_i = 1, \forall i$.

3.1.3 Leaky OR

In practical applications, it is often unfeasible to list all the possible causes of an effect. In this case, we can build a leaky model, as explained in Section 2.3, whose leaky vectorial parameter $P(y_L)$ is given by an single parameter c_L :

$$\begin{cases} P(+y_L) = c_L \\ P(\neg y_L) = 1 - c_L \end{cases} \quad (13)$$

It follows from Equation 8 that when all the explicit causes are absent, i.e., when the configuration of the parents is $\neg\mathbf{x}$, then $\mathbf{X}^+ = \emptyset$ and

$$P(+y|\neg\mathbf{x}) = c_L \quad (14)$$

This means that c_L is the probability that Y is present when all the causes of Y explicit in the model are absent or, put another way, the probability that the implicit causes produce Y . The noisy OR is a particular case of the leaky OR in which $c_L = 0$.

Henrion’s parameters vs. Díez’s parameters The leaky OR, introduced by Henrion [4], is also a particular of the leaky MAX proposed by Díez [1]. However these authors use slightly different parameters to define this model.

In the noisy OR, c_i is the probability of $+y$ when X_i is present and the other X_j ’s are absent. However, in the leaky OR, this probability $P(+y|+x_i, \neg x_{j(j\neq i)})$ corresponds to a different parameter p_i :

$$p_i = P(+y|+x_i, \neg x_{j(j\neq i)}) = c_i + (1 - c_i) \cdot c_L \quad (15)$$

It immediately follows that

$$1 - c_i = \frac{1 - p_i}{1 - c_L} \quad (16)$$

Since $c_L > 0$ in the proper leaky OR, then $p_i > c_i$. This result was foreseeable from the way the noisy OR was built. In fact, we departed from a hypothetical extended model in which the causes of Y were \mathbf{X}^\dagger ; c_i is the probability of $+y$ when X_i is present and all the other causes of Y are absent. In contrast, the definition of p_i requires that all the causes of Y different from X_i and **explicit in the leaky model** ($\mathbf{X} \subset \mathbf{X}^\dagger$) are absent, but admits the possibility that the causes implicit in the leaky model ($\mathbf{X}_I = \mathbf{X}^\dagger \setminus \mathbf{X}$) are present. Since c_i excludes all the causes other than X_i , and p_i admits the presence of some causes other than X_i , it is expected that $p_i > c_i$. A similar argument can be used to justify the inequality $p_i > c_L$.

Henrion’s [4] definition of the leaky OR is based on the p_i ’s. In fact, Henrion’s equation for deriving the conditional probability table was, with a slightly different notation,

$$P(+y|\mathbf{x}) = 1 - (1 - c_L) \cdot \prod_{X_i \in \mathbf{X}^+} \frac{1 - p_i}{1 - c_L} \quad (17)$$

In contrast, the leaky OR considered as a special case of Díez’s [1] leaky MAX for binary variables, is based on the c_i ’s.

Which parameters are more appropriate from the point of view of knowledge engineering? It depends on the knowledge source. When the probabilities are obtained from a database, c_L can be estimated as the proportion of cases in which Y is present and all the causes of Y stored in the database are absent (cf. Eq. 14); if $c_L > 0$ then there must be other causes of Y not recorded in the database. The proportion of “cases in which X_i and Y are present and the X_j ’s recorded in the database take on the value **absent**”

is an estimate of p_i (cf. Eq. 15). In this situation it is impossible to estimate the c_i 's directly from the database, because, as mentioned above, c_i is the probability that Y is present when X_i is present and all the other causes of Y , either included in the database or not, are absent; but obviously we cannot know the value of a variable not recorded in the database.

In contrast, when the probabilities are estimated by human experts, the question that the knowledge engineer should ask in order to obtain Díez's c_i is: "What is the probability that X_i produces Y ?"; in turn, the question for obtaining Henrion's p_i is: "What is the probability Y is present when X_i is present and none of the other causes we are considering in our model is present?". The estimation required by the first can be based on an analysis of the causal mechanism $X_i \rightarrow Y$ and its possible inhibitors (cf. the paragraph preceding Eq. 11). In contrast, the estimation required by the second question should be based on the "statistical" data stored in the expert's memory. Recent work on the elicitation of probabilities for a medical expert system has shown that it is much easier for human experts to answer the first type of questions; this work is coherent with our conjecture that human's estimation of probabilities relies on knowledge of the causal mechanisms and not only on observed frequencies.

In summary, the frequencies we observe in a database correspond to Henrion's p_i 's. In contrast, when the probabilities are estimated by a human expert, we recommend to aim directly at the c_i 's because the corresponding questions are much more brief and intuitive. Fortunately, if c_L is small—as it shall be, because an accurate model must explicitly include all the frequent causes of every anomaly—the difference between each c_i and p_i will be much smaller than the error of the subjective estimate, and the knowledge engineer does not need to care about which parameters he is eliciting from the expert.

3.2 MAX models

The MAX model assumes that there are different causes $\{X_1, \dots, X_n\}$ leading to a certain effect Y that might take on several degrees of severity; it also assumes that the degree reached by Y is the maximum of the degrees produced by each cause if they were acting independently. Henceforth, if the effects accumulate (for example, in economy the effects of different factors tend to accumulate with one another), the MAX model would not be valid.

3.2.1 Deterministic MAX

The deterministic MAX requires that Y and its parents are ordinal variables all sharing the same domain; the value taken on by Y is:

$$y = f_{\text{MAX}}(\mathbf{x}) = \max(x_1, \dots, x_n) \quad (18)$$

3.2.2 Noisy MAX

The noisy MAX can be described by introducing n auxiliary variables $\{Z_1, \dots, Z_n\}$, as discussed in Section 2.2. The relation between Y and the Z_i 's is given by a deterministic MAX, which implies that the domain of every Z_i must be the same as that of Y , and $y = f_{\text{MAX}}(\mathbf{z})$. However, the noisy MAX allows that $\text{dom}(X_i)$ is different from $\text{dom}(Y)$, $\text{dom}(Z_i)$ and $\text{dom}(X_j)$ for other X_j 's.

The parameters for link $X_i \rightarrow Y$ are

$$c_y^{x_i} = P(Z_i = y|x_i) \quad (19)$$

The model defined this way is a **generalized** version of the noisy MAX. The **standard noisy MAX**, however, imposes additional restrictions that lead to an interpretation of the X_i 's as causes that can produce Y . In this case, Y represents an anomaly; the minimum value of Y is termed **normality state** and denoted by $\neg y$, because it corresponds to the absence of anomaly Y . Each X_i has also a **normality state** $\neg x_i$, such that

$$c_{\neg y}^{\neg x_i} = P(Z_i = \neg y|\neg x_i) = 1 \quad (20)$$

In the standard noisy MAX, $Z_i = y$ represents the fact that cause X_i has raised the value of Y from $\neg y$ to y . As a consequence, $c_y^{x_i}$ represents the probability that X_i , when taking on the value x_i , raises Y to a value y .

If all the variables involved in a standard noisy MAX are binary variables of type present/absent or positive/negative and the normality state of every variable corresponds to absent or negative, we obtain the noisy OR.

The noisy MAX was first used by Henrion [4]; Díez [1] formalized the model, coined the term ‘‘MAX gate’’, and developed the leaky MAX and the noisy/leaky MIN.

3.2.3 Leaky MAX

By similarity with the leaky OR, some of the causes of a multivalued effect Y may be left implicit in the model. Under the conditions discussed in Section 2.3, some causes can be implicitly represented by a leaky vector c_y^L , which gives the probability that $Y = y$ when the causes explicit in the model are known to be absent:

$$c_y^L = P(y|\neg \mathbf{x}) \quad (21)$$

The noisy MAX and the leaky OR are particular cases of the leaky MAX.

4 Conclusions

In this paper we have introduced a framework that considers three categories of canonical models: deterministic, noisy and leaky. Each noisy model is based on its deterministic counterpart by introducing an auxiliary variable Z_i for each parent X_i and, optionally, by imposing some restrictions on the values of the conditional probabilities $P(z_i|x_i)$, as

shown in Section 2.2. Analogously, each leaky model is based on its noisy counterpart by grouping the effect of the non-explicit causes into a leaky parameter (a number or a vector). One of the contributions of this paper is the analysis of the conditions required to integrate a leaky model into a larger network. Such conditions should be checked by knowledge engineers when building Bayesian networks.

The general framework offers a unified description of existing models, such as the OR, MAX, AND, and MIN families, as well as Srinivas model [8], and can be used to design new models for specific needs, such as noisy or leaky arithmetic models.

The second contribution of this paper is an analysis of the properties of the OR/MAX family of models—which are those most used in practical applications—with special attention been paid to knowledge engineering issues, such as how to obtain the parameters from a database or which questions the knowledge engineer must pose to the human experts. These and the other models mentioned above are further analyzed in an article on which we are currently working (this paper presented to CAEPIA-2001 is a brief version of that article).

References

- [1] F. J. Díez. Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, pages 99–105, Washington D.C., 1993. Morgan Kaufmann, San Mateo, CA.
- [2] F. J. Díez. Efficient computation for the noisy MAX. Technical Report IA-2001-03, Dpto. Inteligencia Artificial, UNED, Madrid, 2001. A shorter version is included in this volume (CAEPIA-2001).
- [3] I. Good. A causal calculus (I). *British Journal of Philosophy of Science*, 11:305–318, 1961.
- [4] M. Henrion. Some practical issues in constructing belief networks. In L. N. Kanal, T. S. Levitt, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 161–173. Elsevier Science Publishers, Amsterdam, 1989.
- [5] C. Lacave and F. J. Díez. A review of explanation methods for Bayesian networks. To appear in *Knowledge Engineering Review*.
- [6] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.
- [7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988. Revised second printing, 1991.
- [8] S. Srinivas. A generalization of the noisy-OR model. In *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, pages 208–215, Washington D.C., 1993. Morgan Kaufmann, San Mateo, CA.