Approximating Discrete Probability Distributions with Decomposable Models

Francesco M. Malvestuto

Abstract—**A heuristic procedure is presented to approximate an n-dimensional discrete probability distribution with a decomposable model of a given complexity. It is shown that, without loss of generality, the search space can be restricted to a suitable subclass of decomposable models, whose members are called *elementary models*. The selected elementary model is constructed in an incremental manner according to a local-optimality criterion that consists in minimizing a suitable cost function. It is further shown with an example that the solution computed by the procedure is sometimes optimal.**

I. Introduction

It is generally acknowledged that, when designing information systems such as communication, pattern recognition, knowledge and data base systems, the problem that typically arises is that of approximating (or estimating) an n-dimensional discrete probability distribution from a finite number of given marginals (or samples), and storing the distributions in a certain limited amount of machine memory.

In [16] the problem of approximating an n-dimensional discrete probability distribution with a product form was considered. In [5], a method for optimally approximating an n-dimensional discrete probability distribution with a set of $n-1$ first-order dependence relationships among the n variables (i.e., bidimensional marginal distributions) was presented. The procedure in [5], which is inspired by a hill-climbing search strategy, involves an optimization process to construct a dependence tree of maximum weight. As a matter of fact, in many applications unidimensional and bidimensional marginal distributions are inappropriate to extract features that distinguish different patterns.

Better approximations can be obtained by considering *interaction models* [2], [7] (also called *dependency models* [23]) made up of arbitrary sets of marginal distributions. *Decomposable models* [2], [7], [10], [11] are special interaction models that possess a number of desirable properties. The two following properties stress the desirability of decomposable models both from a semantic and formal point of view:

(semantic property) a decomposable model can be represented by a "Markovian belief network" [23],

(formal property) the approximation generated by a decomposable model has a "product form" [2], [10], [11], [19].

In this paper, we address the problem of finding the model that generates the best approximation to a given discrete probability distribution, selected among the decomposable models of a given complexity.

Apart from special cases, this problem can be solved exactly only by exhaustive search [23]. Now, when the number of variables involved is high, exhaustive search is impractical since the number of models that must be examined is enormous. Therefore, one is forced to relax the optimality requirement and settle for finding a good solution using only a reasonable amount of search effort.

Manuscript received January 2, 1990; revised August 5, 1990.

The author was with the Studi Department, ENEA, Via Regina Margherita 125, 00198 Rome, Italy. He is now with the Electrical Engineering Department, Universita Degli Studi di L'Aquila, 67040 Poggio di Roio, L'Aquila, Italy.

IEEE Log Number 9040002.

Some procedures proposed in the analysis of contingency tables [12], [29] might be applied, but have the disadvantage that, since they adopt a backward-selection strategy, when the number of the involved variables is high, the solution is found after evaluating a large set of intermediate solutions. In order to reduce the search effort, in this paper we propose a heuristic procedure that directly yields a suboptimal solution, based on the "hill climbing" search strategy, which, though the simplest and cheapest of search strategies, nevertheless proves to be successful in finding optimal solutions when the search space is restricted to tree-dependence approximations [5].

II. Background

Let $X = \{A_1, \cdots, A_n\}$ be a set of discrete random variables, whose value sets will be denoted by $range(A_1), \cdots, range(A_n)$. Given a joint probability distribution $p(x)$, an approximation to $p(x)$ can be obtained by hypothesizing a structural model [2] over X, that is, a specific stochastic relation among the variables in X. In what follows, by *domain* of a structural model μ we mean the set of all probability distributions over X that are perfectly fitted by the model μ.

The simplest example of structural model over X is given by an *independence model*, which consists in viewing the variables in X partitioned into (stochastically) independent sets. Let $M = \{X_1, \cdots, X_m\}$ be such a partition of X and μ the independence model generated by M. The domain of μ contains all the probability distributions $p(x)$ that can be factorized as

$$p(x) = p_1(x_1) \cdots p_m(x_m)$$

where $p_h(x_h)$ is the marginal of $p(x)$ relative to $X_h (h = 1, \cdots, m)$. In information-theoretic terms, independence of $X_1, \cdots, X_m$ is expressed by the condition that the (*mutual*) *information* [26] of $X_1, \ldots, X_m$, defined as

$$I(X_1, \cdots, X_m) = H(X_1) + \cdots + H(X_m) - H(X)$$

is zero. Here, $H(X_h)$ and $H(X)$ stand for the entropies [26] of the probability distributions $p_h(x_h)$ and $p(x)$, that is,

$$H(X_h) = -\sum\nolimits_{x_h} p_h(x_h) \log p_h(x_h)$$

and

$$H(X) = -\sum\nolimits_{x} p(x) \log p(x).$$

A widely used structural model, which is a natural extension of an independence model, is a *conditional-independence model* [8], [12] (or *zero-partial-association model* [29]), which consists in viewing the variables in X partitioned into nonempty sets $Y_0, Y_1, \cdots, Y_m$ such that $Y_1, \cdots, Y_m$ are conditionally independent given Y_0. The domain of such a model contains all the probability distributions $p(x)$ that can be factorized as follows

$$p(x) = p_1(y_0, y_1) \cdots p_m(y_0, y_m) / p_0(y_0)^{m-1}.$$

In information-theoretic terms, the conditional independence of $Y_1, \cdots, Y_m$ given Y_0 is expressed by the condition that the *average conditional information* [26] of $Y_1, \cdots, Y_m$ given Y_0, defined as

$$I(Y_1, \cdots, Y_m / Y_0) = H(Y_1/Y_0) + \cdots + H(Y_m/Y_0) - H(Y_1 \cup \cdots \cup Y_m / Y_0)$$

is zero. Here, $H(Z/Y)$ denotes the *average conditional entropy* [26] of the variable set Z given the variable set Y, defined as $H(Y \cup Z) - H(Y)$.

The most general form of independence is taken into account by an *interaction model* over X, which groups the variables in X into (possibly overlapping) sets (to be called the *generators* of the model) having "interaction zero" [13], [18]. If $X_1, \cdots, X_m$ denote the generators of such an interaction model μ, as usual we assume that 1) $X = \cup_{h=1,\cdots,m} X_h$ and that 2) no generator of μ is a subset of any other. According to the terminology used in graph theory [1], [15], the set $M = \{X_1, \cdots, X_m\}$ will be called the *generating hypergraph* (also called "generating class" [2], [7], [8], [9], [10]) of μ. The domain P_μ of the interaction model μ generated by the hypergraph M contains all the probability distributions $p(x)$ that can be expressed in the following multiplicative form

$$p(x) = f_1(x_1) \cdots f_m(x_m) \tag{1}$$

where $f_1(x_1) \cdots f_m(x_m)$ are implicit functions uniquely determined by the condition that the multiplicative form (1) represent an *extension* of the marginals $p_1(x_1), \cdots, p_m(x_m)$ of $p(x)$, that is, for all $h = 1, \cdots, m$

$$\begin{aligned} p_h(x_h) &= \sum_{A \in X - X_h} \sum_{a \in \text{range}(A)} p(x) \\ &= f_h(x_h) \left[\sum_{A \in X - X_h} \sum_{a \in \text{range}(A)} \right] \prod_{j \neq h} f_j(x_j). \end{aligned}$$

An interaction model μ over X can be represented by an ordinary graph, to be called a *dependence graph* denoted by G_μ, which has X as its node set and an edge between every pair of variables that are contained in a single generator of μ. An expressive property of G_μ is that if Y_0 is a cutset and $Y_1, Y_2, \cdots, Y_m$ are the connected components of the graph obtained from G_μ by deleting Y_0, then $Y_1, Y_2, \cdots, Y_m$ are conditionally independent given Y_0 [8].

Consider now the problem of approximating a given probability distribution $p(x)$ using interaction models. Let μ be the interaction model generated by the hypergraph $M = \{X_1, \cdots, X_m\}$, the approximation $p_\mu(x)$ to $p(x)$ based on μ is defined as the extension of the marginals $p_1(x_1), \cdots, p_m(x_m)$ of $p(x)$, which is uniquely determined by the membership in P_μ, that is, by the multiplicative form (1) [14]. It is well-known [4], [16] that $p_\mu(x)$ is the maximum-entropy extension of the marginals $p_1(x_1), \cdots, p_m(x_m)$ of $p(x)$. Moreover, $p_\mu(x)$ can be computed using an iterative proportional fitting procedure [4], [21].

Now, let us consider the problem of selecting the "best" approximation to an n-dimensional probability distribution $p(x)$ among the approximations based on interaction models of a given complexity. The aspects of this problem with regards to complexity will be discussed in the next section, while the remainder of this section will be devoted to the criterion for evaluating the accuracy of the approximation $p_\mu(x)$ to $p(x)$ based on an arbitrary interaction model μ . As usual [4], [5], [14], [16], this will be measured by the *information divergence* [6] (or discrimination information [13], [14] or cross-entropy [28] or Kullback–Leibler distance [3])

$$\begin{aligned} D(p, p_\mu) &= \sum_x p(x) \log [p(x)/p_\mu(x)] \\ &= -H(X) - \sum_x p(x) \log p_\mu(x) \end{aligned} \tag{2}$$

which can be interpreted as the difference of the information contained in $p(x)$ and that contained in $p_\mu(x)$ about $p(x)$ [16]. The information divergence $D(p, p_\mu)$ is a non-negative quantity that vanishes if and only if $p_\mu(x)$ coincides with $p(x)$, that is, if and only if $p(x)$ belongs to the domain of the interaction model μ.

According to the choice of measuring the accuracy of approximation with the information divergence, given any two models μ and μ' over X, μ' provides a *better* solution than μ to the problem of approximating $p(x)$ if $D(p, p_{\mu'}) \leq D(p, p_\mu)$. Of course, the information divergence is not the only way of measuring the closeness of $p_\mu(x)$ to $p(x)$; an alternative might be the so-called χ^2 distance [3] defined as

$$\chi^2(p, p_\mu) = \sum_x [p(x) - p_\mu(x)]^2 / p(x)$$

but, the χ^2 distance has not some desirable properties (e.g., additive property) that are possessed by the information divergence. On the other hand, it is well-known [3] that $D(p, p_\mu) \approx (1/2)\chi^2(p, p_\mu)$, so that minimizing $\chi^2(p, p_\mu)$ would lead to the same results that are obtained by minimizing $D(p, p_\mu)$.

Furthermore, recently it has been proved that under certain assumptions the minimization of the information divergence can be derived by minimizing an upper bound of the Bayes error rate [30]. Other properties of the minimum information divergence principle are studied from an axiomatic point of view in [28].

III. The Search Space

In this section, we will specify the complexity of the interaction models among which the solution to the approximation problem will be searched for. We shall do this by introducing the notion of a decomposable model of rank k, which can be regarded as a natural extension of a first-order dependence tree used by Chow and Liu in [5].

We begin by considering the set of all possible interaction models over a given set X of n random variables. This set can be partially ordered according to the relationship of "refinement": μ is *finer* than μ' (or, equivalently, μ' is *coarser* than μ) if each generator of μ is a subset of some generator of μ'. It is obvious that if μ is finer than μ', then the domain of μ' is a subset of the domain of μ.

The infimum ν and the supremum ϵ of the partial ordering induced by refinement are generated respectively by the *point hypergraph* X and the *trivial hypergraph* $\{X\}$. The model ν assumes independence among the n variables in X and the model ϵ makes no assumptions.

The search space, in which we set out in pursuit of the solution to the problem of approximating $p(x)$, is the class of all *decomposable models* of *rank* k, which are defined as interaction models over X satisfying the two following constraints.

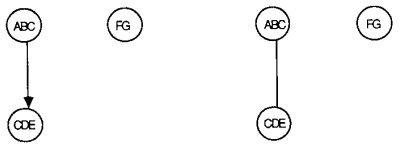
A. Decomposable Models

An interaction model μ is *decomposable* [7] (also called *multiplicative* in [29]) if an ordering $(X_1, \cdots, X_m)$ of its generators exists such that for each $h > 1$ either

1) $(X_1 \cup \cdots \cup X_{h-1}) \cap X_h = \emptyset$, or
2) $(X_1 \cup \cdots \cup X_{h-1}) \cap X_h \neq \emptyset$ and there is an integer $j(h) < h$ such that $(X_1 \cup \cdots \cup X_{h-1}) \cap X_h = X_{j(h)} \cap X_h$.

Such an ordering, to be called a *running intersection ordering* (RIO), can be pictured by a collection T of *arborescences* (directed trees) [1] having as its vertex set the generating hypergraph of μ; the roots of the arborescences in T are X_1 and all the generators of type 1), and every arc in T is of the form $(X_{j(h)}, X_h)$. X_h being a generator of type 2). The undirected graph underlying T is called a *join forest* [17], [23], and each traversal of such a join forest starting at any node determines an appropriate RIO of the generators of μ.

In what follows, given a decomposable model and an RIO $(X_1, \cdots, X_m)$. X_1 as well as the generators of type 1) will be called *root generators*, and the generators of type 2) will be called *descending generators*.



(a) (b)

Fig. 1. (a) Graph representation of the RIO (ABC, CDE, FG). (b) Join forest associated to the hypergraph $M = \{ABC, CDE, FG\}$.

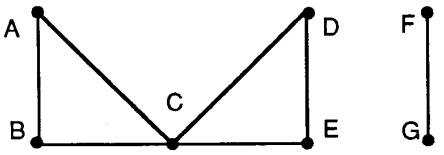


Fig. 2. Dependence graph associated to the model generated by the hypergraph $\{ABC, CDE, FG\}$.

Example: Let us consider the model generated by the hypergraph $M = \{ABC, CDE, FG\}$. M is decomposable for (ABC, CDE, FG) is an RIO , with respect to which ABC and FG are root generators, and CDE is a descending generator (see Fig. 1).

Notice that the model ϵ generated by the trivial hypergraph is decomposable.

The constraint of decomposability is here introduced because decomposable models possess a number of desirable properties [2], [9], [10], [11], [17], [19]. In particular, one has that does the following

1) *Decomposable models are characterized by chordal, conformal dependence graphs* [8], [17], [23].
   We require some preliminary notions of graph theory. A clique in a dependence graph G_μ is a set of nodes (variables) such that every pair forms an edge of G_μ. A dependence graph G_μ is conformal if every clique is contained in some generator of m. A dependence graph G_m is chordal (or triangulated) if every cycle with at least four distinct nodes has a chord, that is, an edge connecting two non-consecutive nodes of the cycle.
   Notice that by the dependence graph associated to a decomposable model of rank 2 is a forest of trees, called *first-order dependence trees* in [5].
2) *A decomposable model can be interpreted in terms of conditional independence* [8].
   Since the relations of conditional independence can be treated in an axiomatic way [22], [23] and the associated formal system can be used as the inference engine of a common sense logic for reasoning about relevance relations [24], [25], decomposability is a desirable quality of belief networks [23].
3) *The approximation based on a decomposable model has a closed product-form expression* [2], [19], [21].
   Let μ be a decomposable model and $(X_1, \cdots, X_m)$ an RIO of its generators. Let us partition the index vertex $J = \{1, 2, \cdots, m\}$ into the two subsets $J' = \{h \in J : X_h$ is a root generator$\}$ and $J'' = \{h \in J : X_h$ is a descending generator$\}$. Moreover, if X_h is a descending generator and $X_{j(h)}$ is its parent, let us denote by Y_h the nonempty intersection of X_h and $X_{j(h)}$, that is, $Y_h = X_h \cap X_{j(h)}, (h \in J'')$. Then, the approximation $p_\mu(x)$ to a given distribution $p(x)$ can be written as follows

$$p_\mu(x) = \prod_{h \in J'} p_h(x_h) \prod_{h \in J''} p_h(x_h / y_h) \tag{3}$$

or, equivalently,

$$p_\mu(x) = \prod_{h \in J} p_h(x_h) / \left[\prod_{h \in J''} p_h(y_h)\right] \tag{4}$$

As a consequence of (4), the entropy $H_\mu(X)$ of $p_\mu(x)$ amounts to

$$\begin{aligned} H_\mu(X) &= -\sum_x p_\mu(x) \log p_\mu(x) \\ &= -\sum_x p_\mu(x) \Big[\sum_{h \in J} \log p_h(x_h) \\ &\quad - \sum_{h \in J''} \log p_h(y_h)\Big] \\ &= \sum_{h \in J} \Big[-\sum_x p_\mu(x) \log p_h(x_h)\Big] \\ &\quad - \sum_{h \in J''} \Big[-\sum_x p_\mu(x) \log p_h(y_h)\Big] \\ &= \sum_{h \in J} \Big[-\sum_{x_h} p_\mu(x_h) \log p_h(x_h)\Big] \\ &\quad - \sum_{h \in J''} \Big[-\sum_{y_h} p_h(y_h) \log p_h(y_h)\Big]. \end{aligned}$$

If we denote the entropy of $p_h(x_h)$ by $H(X_h)$ and the entropy of $p_h(y_h)$ by $H(Y_h)$, the following formula holds

$$H_\mu(X) = \sum_{h \in J} H(X_h) - \sum_{h \in J''} H(Y_h). \tag{5}$$

Example (continued): The dependence graph of the decomposable model μ generated by $M = \{ABC, CDE, FG\}$ is shown in Fig. 2. By inspecting this dependence graph, it is easy to recognize that μ is equivalent to the assumption of the two following relations of independence:

1) $ABCDE$ and FG are independent;
2) AB and DE are conditionally independent given C.

The approximation $p_\mu(abcdefg)$ to a given probability distribution $p(abcdefg)$ based on μ, by (4) is

$$p_\mu(abcdefg) = p(abc)p(cde)p(fg)/p(c).$$

Notice that both the numerator's and the denominator's factors are in a one-to-one correspondence with the nodes and edges of the join forest, shown in Fig. 1(b), respectively. Moreover, by (5) the entropy of $p_\mu(abcdefg)$ is

$$H_\mu(ABCDEFG) = H(ABC) + H(CDE) + H(FG) - H(C).$$

Bounded-Rank Models

A model μ is *of rank* $k (1 < k < n)$ if k is the maximum cardinality of the generators of μ.

It should be noted that such a constraint must be introduced to avoid the trivial solution represented by the model ϵ generated by the trivial hypergraph.

At this point, our approximation problem can be stated as follows.

1) A Minimization Problem: Given a probability distribution $p(x)$ over a set of n discrete random variables, find the approximation $p_\mu(x)$ to $p(x)$ such that

$$D(p, p_\mu) = \min$$

over all decomposable models of rank k over X.

The requirement of decomposability for the interaction models in the search space entails some important simplifications to the approximation problem.

From property (3) the following lemma follows.

Lemma: For every approximation $p_\mu(x)$ to $p(x)$ based on a decomposable model μ, the information divergence $D(p, p_\mu)$ equals the difference between the entropy $H_\mu(X)$ of $p_\mu(x)$ and the entropy $H(X)$ of $p(x)$.

Proof: The term $-\Sigma_x\, p(x) \log\, p_\mu(x)$ in (2) can be expanded as follows

$$\begin{aligned}
-\sum\nolimits_x p(x)\log p_\mu &= -\sum\nolimits_x p(x)\Big[\sum\nolimits_{h\in J}\log p_h(x_h) \\
&\quad -\sum\nolimits_{h\in J''}\log p_h(y_h)\Big] \\
&= \sum\nolimits_{h\in J}\Big[-\sum\nolimits_x p(x)\log p_h(x_h)\Big] \\
&\quad -\sum\nolimits_{h\in J''}\Big[-\sum\nolimits_x p(x)\log p_h(y_h)\Big] \\
&= \sum\nolimits_{h\in J}\Big[-\sum\nolimits_{x_h} p_h(x_h)\log p_h(x_h)\Big] \\
&\quad -\sum\nolimits_{h\in J''}\Big[-\sum\nolimits_{y_h} p_h(y_h)\log p_h(y_h)\Big] \\
&= \sum\nolimits_{h\in J} H(X_h) - \sum\nolimits_{h\in J''} H(Y_h)
\end{aligned}$$

which, by (5), is nothing more than the entropy $H_\mu(X)$ of $p_\mu(x)$. Q.E.D.

As a consequence of this lemma, one has that, since the entropy $H(X)$ of $p(x)$ is independent of the decomposable model μ, minimizing the information divergence $D(p, p_\mu)$ over decomposable models of rank k is equivalent to minimizing the entropy $H_\mu(X)$. By virtue of this result, the following property can be stated.

Theorem 1: Let μ and μ' be two decomposable models over X. If μ' is coarser than μ, then μ' provides a better solution than μ to the problem of approximating any probability distribution over X.

Proof: Let μ and μ' be two decomposable models over X. If μ' is coarser than μ, then $P_\mu \supseteq P_{\mu'}$ and, consequently, $H_\mu(X) \geq H_{\mu'}(X)$. Then, regardless of the probability distribution $p(x)$, from Lemma it follows that $D(p, p_\mu) \geq D(p, p_{\mu'})$. Q.E.D.

Furthermore, since, for every decomposable model μ of rank k, $H_\mu(X)$ depends on $p(x)$ only through its marginals $p_1(x_1), p_2(x_2), \cdots, p_m(x_m)$, the solution of our minimization problem does not require knowledge of the entire distribution $p(x)$; the k-dimensional marginals are sufficient for this purpose. This property proves to be useful in the analysis of contingency tables [2] and for the management of statistical databases [20].

Generally speaking, our minimization problem can be solved in an exact way only by exhaustive search [23]. However, it was proved by Chow and Liu [5] that, if the model search is carried out among dependence trees (i.e., decomposable models of rank 2 generated by connected hypegraphs), the exact solution can be found by using a Kruskal-like step-by-step search procedure. Similarly, optimal decomposable models of rank $k = n - 1$ can be found directly, by ranking the $n(n-1)/2$ average conditional informations of the form

$$\begin{aligned}
&I(A, B/X - \{A, B\}) \\
&\quad = H(X - \{A\}) + H(X - \{B\}) - H(X) - H(X - \{A, B\})
\end{aligned}$$

for each pair A and B of variables in X. Greedy algorithms, which for a given value of k compute a good solution to our minimization problem, are described in Section V.

IV. Uniform and Elementary Models

In this section, we will introduce a special class of decomposable models of rank k, to be called *elementary models* of rank k, and will show that, without loss of generality, the search space of our minimization problem can be restricted to such models. This will be done in two steps. First, we shall prove that, without loss of generality, the search space can be restricted to the class of *uniform decomposable models* of rank k, i.e., decomposable models whose generators are k-sets in the sense that each generator contains exactly k variables. Subsequently, we shall prove that the search space can be further restricted to the class of elementary decomposable models of rank k (whose definition will be introduced later on).

Theorem 2: For every decomposable model μ of rank k, a uniform model of rank k that is coarser than μ always exists.

Proof: We prove the theorem by indicating a constructive method for obtaining a uniform model of rank k from an arbitrary decomposable model μ of rank k. Let $(X_1, X_2, \cdots, X_m)$ be an RIO associated to μ such that $|X_1| = k$. Since at least one generator of μ contains k variables, such an RIO always exists and can be obtained by visiting the nodes of a join tree associated to μ starting at X_1. For $h = 2, 3, \cdots, m$, apply the following procedure to all generators X_h of μ containing less than k variables.

Procedure: Let X_h be a generator containing $k'(k' < k)$ variables. Two cases are to be distinguished depending on whether X_h is a descending generator or a root generator. In the former case, let $X_{j(h)}$ be the parent of X_h and $|X_h - X_{j(h)}| = r$. Then, replace X_h by $X'_h = X_h \cup Y$ where Y is such that:

1) $X_{j(h)} \supset Y \supset X_h \cap X_{j(h)}$;
2) $|Y| = k - r$.

(A set such as Y always exists because $j(h) < h$ and, therefore, $X_{j(h)} = k$). In the latter case, replace X_h by $X'_h = X_h \cup Y$ where Y is any subset of X_{h-1} with exactly $k - k'$ variables.

The result of the replacement of X_h by X'_h in the original RIO is again an RIO because:

1) if X_h is a descending generator with parent $X_{j(h)}$, then

$$(X_1 \cup \cdots \cup X_{h-1}) \cap X'_h = X_{j(h)} \cap X'_h$$

(Notice that X'_h has the same parent that X_h had in the original RIO)

2) if X_h is a root generator, then

$$(X_1 \cup \cdots \cup X_{h-1}) \cap X'_h = X_{h-1} \cap X'_h$$

(Notice that in the new RIO X'_h is a descending generator and X_{h-1} is its parent).

Therefore, after applying this procedure to all generators of μ containing less than k variables, one obtains a uniform model of rank k, that is coarser than the decomposable model μ. Q.E.D.

Example (continued): The decomposable model generated by $M = \{ABC, CDE, FG\}$ is of rank 3 but is not uniform. $N = \{ABC, CDE, EFG\}$ is an example of a uniform decomposable model of rank 3 that can be obtained by applying the previous procedure.

A further restriction of the search space without loss of generality can be achieved after introducing the notion of an *elementary model*, which can be viewed as a generalization of a dependence tree.

Let μ be a uniform decomposable model of rank k and $(X_1, \cdots, X_m)$ is an RIO. Generator X_h $(h > 1)$ is said to be *elementary* if the set $X_h - (X_1 \cup \cdots \cup X_{h-1})$ is a singleton; otherwise, it is said to be *multiple*.

A uniform decomposable model of rank k is an *elementary model of rank* k if an RIO exists in which each X_h $(h > 1)$ is elementary. Thus, an elementary model of rank k has exactly

$m = n - k + 1$ generators and every join forest associated to its generating hypergraph is a tree. On account of this, formula (5) becomes

$$H_\mu(X) = H(X_1) + \sum_{h=2,\cdots,n-k+1} [H(X_h) - H(Y_h)] \quad (6)$$

It should be noted that for $k = n - 1$, an elementary model consists of a pair of conditionally independent variables given the $(n-2)$ remaining variables. For this reason, such models are called "elementary conditional-independence models" [12] (or "elementary zero partial-association models" [29]).

Elementary models of rank k are important because of the following property.

Theorem 3: For every uniform decomposable model μ of rank k, an elementary model of rank k that is coarser than μ always exists.

Proof: We prove the theorem by indicating a constructive method for obtaining such a model starting from an arbitrary RIO $(X_1, X_2, \cdots, X_m)$ of the generators of μ . For $h = 2, 3, \cdots, m$, apply the following procedure to all multiple generators X_h of μ.

Procedure: Insert between X_{h-1} and X_h in the RIO a sequence of k-sets obtained as follows. Two cases are to be distinguished depending on whether X_h is a descending generator or a root generator.

In the former case, if $X_{j(h)}$ is the parent of X_h and $X_h - X_{j(h)} = \{A_1, A_2, \cdots, A_r\} (r > 1)$, then set $Z_i = Y^{(i)} \cup \{A_1, A_2, \cdots, A_i\} (i = 1, \cdots, r-1)$ where $Y^{(1)}, Y^{(2)}, \cdots, Y^{(r-1)}$ is a sequence of subsets of $X_{j(h)}$ such that:

1) $X_{j(h)} \supset Y^{(1)} \supset Y^{(2)} \supset \cdots \supset Y^{(r-1)} \supset X_h \cap X_{j(h)}$;
2) $|Y^{(i)}| = k - i \ (i = 1, \cdots, r-1)$.

Finally, insert between X_{h-1} and X_h in the RIO the sequence $(Z_1, Z_2, \cdots, Z_{r-1})$.

In the latter case, if $X_h = \{A_1, A_2, \cdots, A_k\}$, then set $Z_i = Y^{(i)} \cup \{A_1, A_2, \cdots, A_i\} (i = 1, \cdots, k-1)$ where $Y^{(1)}, Y^{(2)}, \cdots, Y^{(k-1)}$ is a sequence of nonempty subsets of X_{h-1} such that:

1) $X_{h-1} \supset Y^{(1)} \supset Y^{(2)} \supset \cdots \supset Y^{(k-1)}$;
2) $|Y^{(i)}| = k - i \ (i = 1, \cdots, k-1)$.

Finally, insert between X_{h-1} and X_h in the RIO the sequence $(Z_1, Z_2, \cdots, Z_{k-1})$.

One can immediately see that in both cases the result of the insertion of the k-sets Z_i between X_{h-1} and X_h is again an RIO Therefore, after applying this procedure to all multiple generators of m, one obtains an elementary model of rank k that is coarser than the decomposable model μ. Q.E.D.

Example (continued): The hypergraph $N = \{ABC, CDE, EFG\}$ generates a uniform decomposable model of rank 3, which is not elementary. $Q = \{ABC, BCD, CDE, DEF, EFG\}$ is an example of an elementary model of rank 3 that can be obtained by applying the previous procedure.

As a consequence of Theorems 1–3, one has that, without loss of generality, the search space of our minimization problem can be restricted to elementary models of rank k.

V. Greedy Algorithms

In this section we describe a hill-climbing procedure for computing solutions that are, in some sense, locally optimal. When, in Section VI, we shall apply this procedure to a sample distribution, we shall find that the computed solutions are close to optimal, and, in one case, that the solution is actually optimal.

The heuristic technique that will be employed consists in viewing the construction of the generating hypergraph of the goal model as a sequence of $n - k + 1$ decisions, which extend partial solutions (i.e., $\{X_1, \cdots, X_h\}$) by including a single generator, selected in such a way as to preserve decomposability. Each decision is made according to a local-optimality criterion consisting in minimizing a suitable cost function.

Bearing in mind the expression (6) of the objective function $H_\mu(X)$, we define the cost for a partial solution $\{X_1, \cdots, X_h\} (h = 1, \cdots, n-k+1)$ as follows:

for $h = 1$ *cost* $\{X_1\} = H_(X_1)$

for $h = 2, \cdots, m$ *cost* $\{X_1, \cdots, X_h\} = H(X_1) + \Sigma_{i=2,\cdots,h}[H(X_i) - H(Y_i)]$.

Now, for a cost function to work correctly, for all partial solutions $\{X_1, \cdots, X_h\}$ and all extensions $\{X_1, \cdots, X_h, X_{h+1}\}$ we must have [27]:

$$cost\ X_1, \cdots, X_h \leq cost\ \{X_1, \cdots, X_h, X_{h+1}\}.$$

In fact, only if the cost function has this property, can a partial solution $\{X_1, \cdots, X_h\}$ be discarded when its cost is greater than or equal to the cost of a previously computed solution.

However, in our case, when a partial solution $\{X_1, \cdots, X_h\}$ is extended to $\{X_1, \cdots, X_h, X_{h+1}\}$, it might happen that the parents of $X_2, \cdots, X_h$ change, and this would cause an undesirable, nonmonotonic behavior in the cost function. The following example illustrates this anomaly.

If, after choosing $X_1 = ABC$ and $X_2 = CDE$, one attempted to enter the set $X_3 = BCD$ as the third generator of the model, one would find that cost $\{X_1, \cdots, X_2\} \geq$ cost $\{X_1, X_2, X_3\}$. In fact one has

$$cost\ \{ABC, CDE\} = H(ABC) + [H(CDE) - H(C)]$$

$$cost\ \{ABC, CDE, BCD\} = H(ABC) + [H(CDE) - H(C)] + [H(BCD) - H(BC)]$$

and the quantity

$$cost\ \{ABC, CDE\} - cost\ \{ABC, CDE, ACD\} = H(BC) + H(CD) - H(C) - H(BCD)$$

is non-negative for it is nothing more than $I(B, D/C)$, i.e., the average conditional information of B and D given C.

To overcome this difficulty, the generators of the goal model will be selected according to the search scheme that binds each partial solution to be itself an elementary model of rank k. In other terms, the generator X_{h+1} to add to the partial solution $(X_1, \cdots, X_h)$ $(h = 1, \cdots, n-k+1)$ will be selected from among the k-sets that are elementary with respect to $(X_1, \cdots, X_h, X_{h+1})$. At this point, we can state our first greedy algorithm, referred to as *Algorithm G* in what follows.

Algorithm G

Step 1: Among all the k-sets, select X_1 so that the entropy $H(X_1)$ be as little as possible.

Step 2: After choosing $X_1, \cdots, X_h$ select X_{h+1} among all the possible elementary k-sets so that, if $X_{j(h+1)}$ is the parent of X_{h+1} and $Y_{h+1} = X_{h+1} \cap X_{j(h+1)}$, then the quantity $H(X_{h+1}) - H(Y_{h+1})$ be as little as possible.

Step 3: Exit when $n - k + 1$ k-sets have been selected.

The search effort e of Algorithm G can be measured by the number of k-sets that have to be evaluated.

In what follows, $S_k(Y)$ denotes the family of the subsets of Y having cardinality k. If $|Y| = s$, then $|S_k(Y)| = {}^sC_k$, where ${}^sC_k = s!/k!(s-k)!$ is the number of combinations of s variables taken k at a time.

Since X_1 is taken from $S_k(X)$ and $|X| = n$, the degrees of freedom in the choice of X_1 are $e_1 = {}^nC_k$.

As to $X_h + 1$ $(h = 1, \cdots, n-k)$, it is obtained by adding one variable (*entering variable*) to a $(k-1)$-set taken from the following family

$$S_{k-1}(X_1) \cup S_{k-1}(X_2) \cup \cdots \cup S_{k-1}(X_h).$$

In order to evaluate the degrees of freedom in the choice of X_h+1, let us begin by noting that, if $X_{j(i)}$ is the parent of $X_i (i = 2, \cdots, h)$, then

$$S_{k-1}(X_i) \cap (S_{k-1}(X_1) \cup \cdots \cup S_{k-1}(X_{i-1})) = S_{k-1}(X_i) \cap S_{k-1}(X_{j(i)}).$$

Therefore, one has

$$\begin{aligned} &|S_{k-1}(X_1) \cup \cdots \cup S_{k-1}(X_{h-1}) \cup S_{k-1}(X_h)| \\ &\quad = |S_{k-1}(X_1) \cup \cdots \cup S_{k-1}(X_{h-1})| \\ &\qquad + |S_{k-1}(X_h)| - |S_{k-1}(X_{j(h)}) \cap S_{k-1}(X_h)| \\ &\quad = |S_{k-1}(X_1)| + [|S_{k-1}(X_2)| - |S_{k-1}(X_{j(2)}) \cap S_{k-1}(X_2)|] \\ &\qquad + \cdots + [|S_{k-1}(X_h)| - |S_{k-1}(X_{j(h)}) \cap S_{k-1}(X_h)|]. \end{aligned}$$

Furthermore, $S_{k-1}(X_i) \cap S_{k-1}(X_{j(i)})$ contains exactly one set, i.e., $X_{j(i)} \cap X_i$, so that one has

$$\begin{aligned} &|S_{k-1}(X_1) \cup \cdots \cup S_{k-1}(X_{h-1}) \cup S_{k-1}(X_h)| \\ &\quad = k + (h-1)(k-1) = h(k-1) + 1. \end{aligned}$$

Now, since the possible entering variables are $(n-k-h+1)$, the degrees of freedom in the choice of X_{h+1} are

$$e_{h+1} = [h(k-1)+1](n-k-h+1).$$

In conclusion, on account of the formulas:

$$\sum_{h=1,\cdots,m} h = m(m+1)/2$$
$$\sum_{h=1,\cdots,m} h^2 = m(m+1)(2m+1)/6$$

we can conclude that the search effort $e = \Sigma_{h=1,\cdots,n-k+1} e_h$ of Algorithm G amounts to

$$e = {}^nC_k + (n-k)(n-k+1)[(k-1)(n-k+2)+3]/6.$$

In order to appreciate the convenience of our heuristic approach, one must compare e with the effort of an exhaustive search, which can be measured by the number of all possible elementary models of rank k. Since an evaluation of this number is difficult, we shall use the following lower bound E:

$$\begin{aligned} E &= \left(\prod_{h=1,\cdots,n-k+1} f_h\right)/(n-k+1)! \\ &= \prod_{h=1,\cdots,n-k} [1 + h(k-1)]n!/[k!(n-k+1)] \end{aligned}$$

obtained by dividing the number of all possible RIO's by the number of permutations of $n-k+1$ objects (generators). Notice that for $k = n-1$, one has $E = {}^nC_2$, which is exactly the number of elementary conditional-independence models.

One can easily see that, if n is high, then E is far greater than e.

A cheaper algorithm than Algorithm G can be worked out if the ranges of the variables in X do not have the same cardinality. To this end, we need the notion of a *minimal-range* k-set.

The *range* of a set Y of random discrete variables is defined by the Cartesian product of the ranges of the variables in Y. Minimal-range k-sets are k-sets whose ranges have the least cardinality. Now, if Y is a k-set whose range has cardinality t, then $H(Y) \leq H_0(Y) = \log t$ and, therefore, the minimum-entropy k-set is likely to be a k-set minimizing $H_0(Y)$, that is, a minimal-range k-sets. In this way, the search effort made to select the hth generator X_h is reduced by a quantity equal to $e_h - e_h^*$, if e_h^* is the number of the minimal-range k-sets, among which X_h is selected. This leads us to state a second greedy algorithm, called *Algorithm* G^*, which is more efficient than Algorithm G and usually gives the same results, as shown in the example discussed in Section VI.

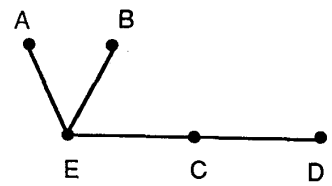


Fig. 3. Dependence graph associated to the model generated by the hypergraph $\{AE, BE, CD, CE\}$.

*Algorithm G**

Step 1: Among all the minimal-range k-sets, select X_1 so that the entropy $H(X_1)$ be as little as possible.

Step 2: After choosing $X_1, \cdots, X_h$ select X_{h+1} among all the minimal-range elementary k-sets so that, if $X_{j(h+1)}$ is the parent of X_{h+1} and $Y_{h+1} = X_{h+1} \cap X_{j(h+1)}$, then the quantity $H(X_{h+1}) - H(Y_{h+1})$ be as little as possible.

Step 3: Exit when $n-k+1$ k-sets have been selected.

VI. Experimental Results

We present experimental results on the application of the proposed procedure to the probability distribution obtained from sample data reported in Table I containing information on the structural habitat of *grahami* and *opalinus* lizards from Whitehouse, Jamaica, taken from Bishop *et al.* [2]. The data consists of observed counts for perch height, perch diameter, insolation, and time-of-day categories for both *grahami* and *opalinus* lizards. The four habitat variables are referred to as variables A, B, C and D, respectively, and species is variable E. The table is of dimension $2 \times 2 \times 2 \times 3 \times 2$.

The following are the values of the entropy for the marginal probability distributions. Moreover $H(ABCDE) = 3.21732$.

1-sets	H	2-sets	H
A	0.64919	AB	1.30243
B	0.66798	AC	1.12634
C	0.47800	AD	1.65839
D	1.01126	AE	1.17158
E	0.54621	BC	1.14281
		BD	1.67593
		BE	1.19777
		CD	1.44673
		CE	1.01847
		DE	1.55183

3-sets	H	4-sets	H
ABC	1.77670	ABCD	2.73327
ABD	2.30811	ABCE	2.27963
ABE	1.81318	ABDE	2.80514
ACD	2.09263	ACDE	2.59497
ACE	1.64171	BCDE	2.62262
ADE	2.17371		
BCD	2.10755		
BCE	1.66783		
BDE	2.19907		
CDE	1.97849		

TABLE I
COUNTS IN STRUCTURAL HABITAT CATEGORIES FOR GRAHAMI AND OPALINUS LIZARDS[a]

Cell (ABCDE)	Observed	Cell (ABCDE)	Observed
11111	20	11112	2
21111	13	21112	0
12111	8	12112	3
22111	6	22112	0
11211	34	11212	11
21211	31	21212	5
12211	17	12212	15
22211	12	22212	1
11121	8	11122	1
21121	8	21122	0
12121	4	12122	1
22121	0	22122	0
11221	69	11222	20
21221	55	21222	4
12221	60	12222	32
22221	21	22222	5
11131	4	11132	4
21131	12	21132	0
12131	5	12132	3
22131	1	22132	1
11231	18	11232	10
21231	13	21232	3
12231	8	12232	8
22231	4	22232	4

[a]From Whitehouse, Jamaica. The sample size is 561.

We applied both Algorithms G and G^* for $k = 2, 3, 4$, and obtained the following results.

Case ($k = 2$). Both Algorithm G and Algorithm G^* select the same model μ, which is generated by the hypergraph $\{AE, BE, CD, CE\}$. Algorithm G does so after selecting

1) $X_1 = CE$ among the ten possible 2-sets;
2) $X_2 = AE$ among the six following 2-sets: *AC, AE, BC, BE, CD, DE*;
3) $X_3 = BE$ among the six following 2-sets: *AB, AD, BC, BE, CD, DE*;
4) $X_4 = CD$ among the four following 2-sets: *AD, BD, CD, DE*.

Algorithm G^* does so after selecting

1) $X_1 = CE$ among the six following 2-sets: *AB, AC, AE, BC, BE, CE;*
2) $X_2 = AE$ among the four following 2-sets: *AC, AE, BC, BE*;
3) $X_3 = BE$ among the three following 2-sets: AB, BC, BE;
4) $X_4 = CD$ among the four following 2-sets: *AD, BD, CD, DE*.

The corresponding value of the cost function amounts to $H_\mu(ABCDE) = 3.26413$. The dependence graph of μ is shown in Fig. 3.

By inspecting this dependence graph, it is easy to recognize that μ is equivalent to the assumption of the two relations of independence:

1) ABE and D are conditionally independent given C;
2) A, B and C are conditionally independent given E.

The exhaustive examination of the one hundred and one possible elementary models of rank 2 shows that μ is the optimal one (!).

Case ($k = 3$). Both Algorithm G and Algorithm G^* select the same model μ, which is generated by the hypergraph $\{ABE, ACE, CDE\}$. Algorithm G does so after selecting

$X_1 = ACE$ among the ten possible 3-sets;
$X_2 = ABE$ among the six following 3-sets: *ABC, ABE, ACD, ADE, BCE, CDE;*

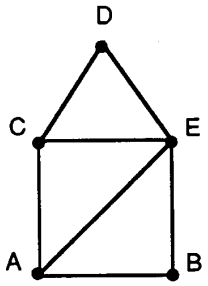


Fig. 4. Dependence graph associated to the model generated by the hypergraph $\{ABE, ACE, CDE\}$.

$X_3 = CDE$ among the five following 3-sets: *ABD, ACD, ADE, BDE, CDE.*

Algorithm G^* does so after selecting

$X_1 = ACE$ among the four following 3-sets: *ABC, ABE, ACE, BCE*;
$X_2 = ABE$ among the three following 3-sets: *ABC, ABE, BCE*;
$X_3 = CDE$ among the five following 3-sets: *ABD, ACD, ADE, BDE, CDE.*

The corresponding value of the cost function amounts to $H_\mu(ABCDE) = 3.24333$. The dependence graph of μ is shown in Fig. 4.

By inspecting this dependence graph, it is easy to recognize that μ is equivalent to the assumption of the two relations of independence:

1) ABE and D are conditionally independent given C;
2) A, B and C are conditionally independent given E.

The exhaustive examination of the 70 possible elementary models of rank 3 shows that there are three models better than μ, namely,

$\mu_1 = \{ABE, BDE, CDE\}$ with $H_{\mu_1}(ABCDE) = 3.24115$ (the optimal model)

$\mu_2 = \{ABE, ADE, CDE\}$ with $H_{\mu_2}(ABCDE) = 3.24197$

$\mu_3 = \{ABE, BCE, CDE\}$ with $H_{\mu_3}(ABCDE) = 3.24327$

Case ($k = 4$). Both Algorithm G and Algorithm G^* select the same model μ, which is generated by the hypergraph $\{ABCE, ACDE\}$. Algorithm G does so after selecting

1) $X_1 = ABCE$ among the five possible 4-sets;
2) $X_2 = ACDE$ among the four remaining 4-sets.

Algorithm G^* does so after selecting

1) $X_1 = ABCE$ as the minimal-range 4-set;
2) $X_2 = ACDE$ among the four remaining 4-sets.

The corresponding value of the cost function amounts to $H_\mu(ABCDE) = 3.23282$. The dependence graph μ is shown in Fig. 5.

By inspecting this dependence graph, it is easy to recognize that μ is equivalent to the assumption of the hypothesis of conditional independence of B and D given ACE.

The exhaustive examination of the ten possible elementary models of rank 4 shows that there are three models better than μ, namely,

$\mu_1 = \{ABCD, ACDE\}$ with $H_{\mu_1}(ABCDE) = 3.22640$ (the optimal model)

$\mu_2 = \{ABDE, BCDE\}$ with $H_{\mu_2}(ABCDE) = 3.22869$

$\mu_3 = \{ABCD, ABDE\}$ with $H_{\mu_3}(ABCDE) = 3.23030$.

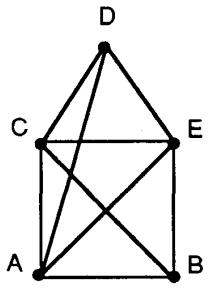


Fig. 5. Dependence graph associated to the model generated by the hypergraph $\{ABCE, ACDE\}$.

The same five-way table was used by Havránek [12], who applied a backward-selection procedure. After evaluating all of the elementary conditional-independence models, of which only six ("baseline models") were accepted, his procedure backwards steps by examining all the decomposable models that are finer than the baseline models. At the end, the two following decomposable models of rank 3 were selected

$$\{ABE, CE, D\} \quad \text{and} \quad \{ABE, BDE, CD\}.$$

Both of them are finer than the model selected by our greedy algorithms, which therefore turns out to be better than both.

VII. Conclusion

We have treated the problem of approximating a given multidimensional probability distribution with interaction models as a minimization problem, which uses information divergence as its cost function and the class of decomposable models of a certain rank as its search space.

The optimal solution to this problem is known to require an exhaustive examination of all models in the search space, except for the case rank=2, which can be solved by using a greedy selection procedure.

We have presented two greedy algorithms, which show that, if one relaxes the optimality requirement and is content with a suboptimal solution, even in the general case (rank$\geq$2) the search effort is bearable. Both of the proposed algorithms restrict the search space to elementary models, which are a subclass of decomposable models. Such a restriction is made on account of the two following facts:

1) the optimal solution to the minimization problem is an elementary model;
2) in the search space of elementary models, the cost function has an additive, closed form.

The additivity of the cost function has been exploited to select a suboptimal solution, which is built up in an incremental manner by using a greedy selection procedure, based on the minimization of the single additive components of the cost function.

Such a selection procedure, when compared with other procedures existing in literature, has two unquestionable advantages:

1) the selected solution is found directly (that is, without passing through intermediate solutions) and, hence, faster;
2) the solution does not require knowledge of the entire distribution to be approximated, but only of its k-dimensional marginals.

Acknowledgment

The author is grateful to an anonymous referee for valuable suggestions, which permitted him to improve Algorithm G^*.

References

[1] C. Berge, *Graphs and Hypergraphs.* Amsterdam, The Netherlands: North-Holland, 1973.
[2] Y. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press, 1975.
[3] A. Borovkov, *Statistique Mathematique.* Moscow, USSR: MIR, 1987.
[4] D.T. Brown, "A note on approximations to probability distributions to reduce storage requirements," *Inform. Contr.*, vol. 2, pp. 386–392, 1959.
[5] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 462–467, 1968.
[6] I. Csisazr, "I-Divergence geometry of probability distributions and minimization problems," in *Ann. Probability*, vol. 3, pp. 146–158, 1975.
[7] J. N. Darroch, "Interaction models," *Encycl. Statist. Sci.*, vol. 4. New York: Wiley, 1983, pp. 182–187.
[8] J. N. Darroch, S. L. Lauritzen, and T. P. Speed, "Markov fields and log-linear interaction models for contingency tables," in *Ann. Statistics* vol. 8, pp. 522–539, 1980.
[9] L. A. Goodman, "The multivariate analysis of qualitative data: Interaction among multiple classifications," *J. Amer. Stat. Ass.* vol. 65, pp. 226–256, 1970.
[10] S. J. Haberman, "The general log-linear model," Ph. D. dissertation, Dep. Statist., Univ. Chicago, 1970.
[11] S.J. Haberman. *The Analysis of Frequency Data.* Chicago, IL: Univ. Chicago Press, 1974.
[12] T. Havránek, "A procedure for model search in multidimensional contingency tables," *Biometrics*, vol. 40, pp. 95-100, 1984.
[13] H. H. Ku and S. Kullback, "Interaction in multidimensional contingency tables: An information theoretic approach," *J. Res. Nat. Bureau Standards–Math. Sci.*, vol. 72B, pp. 159–199, 1968.
[14] ——, "Approximating discrete probability distributions," *IEEE Trans. Inform. Theory*, vol. IT–15, pp. 444–447, 1969.
[15] S. L. Lauritzen, T. P. Speed, and K. Vijayan, "Decomposable graphs and hypergraphs", *J. Austr. Math. Soc. (Series A)*, vol. 36, pp. 12–29, 1984.
[16] P. M. Lewis, "Approximating probability distributions to reduce storage requirements," *Inform. Contr.*, vol. 2, pp. 214–225, 1959.
[17] D. Maier, *The Theory of Relational Databases.* Rockville, MD: Computer Science Press, 1983.
[18] F.M. Malvestuto, "Statistical treatment of the information content of a data base," *Inform. Syst.*, vol. 11, pp. 211–223, 1986.
[19] ——, "Existence of extensions and product extensions for discrete probability distributions," *Discrete Math.*, vol. 69, pp. 61–77, 1988.
[20] ——, "A universal-table model for categorical databases," *Inform. Sci.* vol. 49, pp. 203–223, 1989.
[21] ——, "Computing the maximum-entropy extension of discrete probability distributions," *Comput. Stat. Data Anal.*, vol. 8, pp. 299–311, 1989.
[22] ——, "A unique formal system for binary decompositions of database relations, probability distributions and graphs," *Inform. Sci.*, vol. 52, 1992.
[23] J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* Palo Alto, CA: Morgan Kauffman, 1988.
[24] J. Pearl and A. Paz, "Graphoids: a graph-based logic for reasoning about relevance relations," in *Proc. European Conf. on Artificial Intelligence*, 1986.
[25] J. Pearl and T. Verma, "The logic of representing dependencies by directed graphs," in *Proc. American Ass. for Artificial Intelligence Conf.*, 1987.
[26] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes.* San Francisco, CA: Holden-Day, 1964.
[27] E.M. Reingold, J. Nievergelt, and N. Deo, *Combinatorial Algorithms: Theory and Practice.* Englewood Cliffs, NJ: Prentice-Hall, 1977.
[28] J.E. Shore and R.W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory* vol. 26, pp. 26–37, 1980.
[29] N. Wermuth, "Model search among multiplicative models," *Biometrics*, vol. 32, pp. 253–263, 1976.
[30] S.K.M. Wong and F.C.S. Poon, "Comments on approximating discrete probability distributions with dependence trees," *IEEE Trans. Patt. Anal. Mach. Intell.* vol. 11, pp. 333–335, 1989.