Michael A. Pittarelli

Probabilistic databases and decision problems: Results and a conjecture

Kybernetika, Vol. 29 (1993), No. 2, 149--165

Persistent URL: http://dml.cz/dmlcz/124561

KYBERNETIKA — VOLUME 29 (1993), NUMBER 2, PAGES 149 - 165

PROBABILISTIC DATABASES AND DECISION PROBLEMS: RESULTS AND A CONJECTURE

MICHAEL PITTARELLI

An algebra applicable to recently introduced probabilistic models of data and which resembles the relational algebra is presented. It is shown to support various strategies for decision-making with information in the form of a probabilistic database. A conjecture is stated which, if true, could be exploited to reduce, without loss of information, the cost of solving decision problems involving databases with large numbers of attributes.

1. INTRODUCTION

Recently, a number of closely related probabilistic data models have been introduced [2, 6, 29]. These models include the standard relational model [22] as a special case while permitting the representation of uncertainty regarding the entities being modelled and their interaction in terms of classical probability (vs. non-probabilistic fuzzy measures, in terms of which some extensions of the relational model have been characterized [30]).

One advantage of a probabilistic treatment of database imprecision and uncertainty is that it supports a standard Bayesian approach to decision analysis. In this paper, methods for decision-making with information in the form of a probabilistic database are developed. A conjecture is presented which, if true, could be exploited to greatly reduce the cost of solving decision problems when the available data involve large numbers of irrelevant attributes.

2. BASIC DEFINITIONS AND CONCEPTS

Mathematically, a relational database instance is a (finite) collection of relations on finite domains; i.e., a set of subsets of Cartesian products of finite sets. Each of these finite sets is the *domain* of an *attribute* (or *variable*). Let $\mathrm{dom}(M)$ denote the domain of attribute M. The set of attributes V_i on which an element r_i of a database $R = \{r_1, \ldots, r_m\}$ is defined is its *scheme*, denoted $s(r_i)$. For relation r_i with scheme V_i,

$$r_i \subseteq \bigtimes_{M \in V_i} \mathrm{dom}(M) = \mathrm{dom}(V_i) = \mathrm{dom}(s(r_i))$$

In what follows, a relation r_i will be replaced with its characteristic function. It is a short step from this notion of a relational database to that of a probabilistic database. Instead of a collection $R = \{r_1, \ldots, r_m\}$ with $r_i : \mathrm{dom}(s(r_i)) \longrightarrow \{0, 1\}$, one has a collection $K = \{p_1, \ldots, p_m\}$ with $p_i : \mathrm{dom}(s(p_i)) \longrightarrow [0, 1]$ and $\sum_t p_i(t) = 1$. (A probabilistic database corresponds to a *knowledge store* in Perez and Jiroušek's INES model [26] and to a *structure system* in *reconstructability analysis* [16].)

Let R_V and P_V denote, respectively, the set of all non-empty relations and all probability distributions definable on the scheme V (i.e., on the set of atoms $\mathrm{dom}(V)$). It is shown in [6] that the mapping $t_{rp} : R_V \longrightarrow P_V$, where

$$t_{rp}(r)(t) = r(t) \Big/ \sum_t r(t) ,$$

preserves standard relational data dependencies (functional, join, etc.) in probabilistic form (characterized in terms of conditional and relative entropy). In [29] it is shown that the left inverse

$$t_{pr}(p)(t) = \begin{cases} 1, & \text{if } p(t) > 0, \\ 0, & \text{otherwise} \end{cases}$$

of t_{rp} is a homomorphism from various probabilistic to relational systems defined in terms of standard relational operators and probabilistic analogues. Only three probabilistic operators are discussed in this paper: *projection*, *extension*, and *join*. (For additional operators, see [29].)

For a scheme V, the tuples $t \in \mathrm{dom}(V)$ are viewed as mutually exclusive and exhaustive elementary events. Thus, probabilities $p(S)$, $S \subseteq (V)$, are computed as

$$p(S) = \sum_{t \in S} p(t).$$

The database operation of projection is a special case.

Definition. Let $A \subseteq s(p)$. Then $\pi_A(p)$ denotes the *projection* of distribution p onto A,

$$\pi_A(p)(x) = \sum_{t \in \mathrm{dom}(s(p)),\, t[A]=x} p(t),$$

where, for tuples $w \in \mathrm{dom}(W)$ and $b \in \mathrm{dom}(B)$, $B \subseteq W$, $w[B] = b$ iff w and b agree on all attributes in scheme B.

The term 'projection', rather than 'marginalization', is used for consistency with relational database terminology, since

$$\pi_A(t_{pr}(p)) = t_{pr}(\pi_A(p)),$$

where

$$\pi_A(r)(x) = \max_{t \in \mathrm{dom}(s(r)),\, t[A]=x} r(t).$$

For a set D of distributions, $\pi_V(D)$ denotes the image of D under the mapping π_V. Thus, for any family $(D_i)_{i\in I}$ of sets,

$$\pi_V\left(\bigcup_{i\in I} D_i\right) = \bigcup_{i\in I} \pi_V(D_i) \tag{1}$$

and

$$\pi_V\left(\bigcap_{i\in I} D_i\right) \subseteq \bigcap_{i\in I} \pi_V(D_i). \tag{2}$$

Definition. A *model* of a scheme V is a set $X = \{V_1, \ldots, V_m\}$ such that $V_j \subseteq V$ and $V_i \not\subseteq V_j$ for all $i, j \in \{1, \ldots, m\}$. ($X$ will sometimes be referred to as a model of a distribution with scheme V.)

If X is also a cover of V, i.e., $V = \bigcup_{j=1}^{m} V_j$, then X satisfies the definition of a *reduced hypergraph* over V [3]. Normally, attention is restricted to reduced hypergraph models of a given scheme V.

If X is the set of schemes for the elements of a database K, then K is said to be defined over X, or to have *structure* X. Let $s(K)$ denote the structure of K:

$$s(K) = \{s(p) \mid p \in K\}.$$

Unless stated otherwise, the structure of a database K will always be taken in this paper to be a model, in the above sense, of $\bigcup_{V_i \in s(K)} V_i$. However, structures that are not models are sometimes useful, for example, when a projection $\pi_A(p_i)$ for some $p_i \in K$ is to be accessed frequently.

A distribution p with scheme V may be projected onto a model $X = \{V_1, \ldots, V_m\}$ of V to form a probabilistic database $K = \{p_1, \ldots, p_m\}$, where

$$K = \pi_X(p) = \{\pi_{V_1}(p), \ldots, \pi_{V_m}(p)\}.$$

A database K formed in this way is guaranteed to exhibit various forms of consistency; e.g., for any p_i and p_j in K,

$$\pi_A(p_i) = \pi_A(p_j),$$

for all $A \subseteq s(p_i) \cap s(p_j)$.

Definition. A model X is a *refinement* of model Y (and Y is an *aggregate* or *coarsening* of X), denoted $X \leq Y$, iff for each $V_x \in X$ there exists a $V_y \in Y$ such that $V_x \subseteq V_y$ [5, 9].

Example. $\{\{A\}, \{B,C\}\}$ is a refinement of $\{\{A,B\}, \{B,C\}, \{D\}\}$.
$\{\{A\}, \{B,C\}\}$ is not a refinement of $\{\{B,C\}, \{D,E\}\}$.

The set of all models over V together with the refinement ordering is a lattice [4]. Any pair of models has a greatest lower bound equal to their least refined common refinement and a least upper bound equal to the most refined structure of which they are both refinements. The universal upper bound of the lattice of models over V is $\{V\}$; the lower bound is $\{\emptyset\}$. The reduced hypergraphs also form a lattice, with universal lower bound $\{\{v\} \mid v \in V\}$.

A database K may be projected onto a refinement X of $s(K)$ in the obvious way to form a database

$$\pi_X(K) = \{\pi_{V_x}(p) \mid V_x \in X,\ V_k \in s(K),\ V_x \subseteq V_k,\ p \in K,\ p \in P_{V_k}\}\,.$$

The function symbol 'π' is thus quite overloaded (as it is also in relational database theory). Its precise meaning in any context depends on its parameterization and its argument; but in all cases, marginal probabilities are produced from given joint probabilities.

Definition. The *extension* of a distribution p with scheme A to the scheme V, $A \subseteq V$, is the set of all preimages p' with scheme V of p under the mapping π_A:

$$E^V(p) = \{p' \in P_V \mid \pi_A(p') = p\}\,.$$

The extension of a database K is the intersection of the extensions of its elements:

$$E^V(K) = \bigcap_{p \in K} E^V(p).$$

Thus, $E^V(\pi_X(p))$ is the set of all preimages of the database $\pi_X(p)$ under the mapping π_X; any model X of V partitions P_V into classes $E^V(\pi_X(p))$ equivalent with respect to projections onto X. Any $E^V(p)$ or $E^V(K)$ is a convex polyhedron, the set of solutions to the system of linear equations determined by the projection conditions. (If $s(K)$ is a cover of V, then $E^V(K)$ may be abbreviated $E(K)$.)

Example. The database below represents partial information regarding the contents of a box of wooden blocks.

Color	$p_1(t)$	Shape	$p_2(t)$
Black	0.7	Sphere	0.6
White	0.3	Cube	0.4

Its extension to {Color, Shape} is the set of solutions p to the system

$$\begin{aligned}
&p(\text{Black, Sphere}) + p(\text{Black, Cube}) = 0.7\\
&p(\text{White, Sphere}) + p(\text{White, Cube}) = 0.3\\
&p(\text{Black, Sphere}) + p(\text{White, Sphere}) = 0.6\\
&p(\text{Black, Cube}) + p(\text{White, Cube}) = 0.4\\
&p(t) \geq 0.
\end{aligned}$$

(The equations imply that $\sum_t p(t) = 1$.) From just the information given, it cannot be determined which of the infinitely many members of $E(K)$ is the actual joint distribution over {Color, Shape}.

The smaller the set $E(\pi_X(p))$, the more information regarding p is embodied in its projections onto X. At the extreme, $E(\pi_X(p)) = \{p\}$, and p is said to be *identifiable* from X.

Definition. The *join* of a database $K = \{p_1, \ldots, p_m\}$ is the element $J(K)$ of $E(K)$ with *maximum entropy*:

$$J(K) = \arg \max_{p \in E(K)} -\sum_t p(t) \cdot \ln(p(t)).$$

Since $E(K)$ is convex, $J(K)$ is unique [14]. For a set Q with unique maximum entropy element, let $J(Q)$ denote that element. Determination of $J(Q)$ for arbitrary constraints defining Q is usually a difficult optimization problem. However, this is not the case when the constraints are in the form of a probabilistic database. Projection (i.e., addition), multiplication and division are sufficient.

Definition. Let $K = \{p_1, p_2\}$, with structure $\{V_1, V_2\}$. The *pairwise join* of K is the probability distribution $PJ(K) \in E(K)$ whose components are calculated as

$$PJ(K)(t) = p_1(x) \cdot p_2(y) \Big/ \sum_{z | z[V_1 \cap V_2] = y[V_1 \cap V_2]} p_2(z),$$

where $x = t[V_1]$ and $y = t[V_2]$. (The denominator of the above expression equals 1 if $V_1 \cap V_2 = \emptyset$.)

For sets of variables V_1, V_2, and V_3, V_1 is *conditionally independent* of V_2, given V_3, iff $p(t_{12}|t_3) = p(t_1|t_3) \times p(t_2|t_3)$, for all $t_{12} \in \mathrm{dom}(V_1 \cup V_2)$ and $t_3 \in \mathrm{dom}(V_3)$, where $t_1 = t_{12}[V_1]$ and $t_2 = t_{12}[V_2]$. It follows immediately that for a model $\{V_1, V_2\}$ of p with scheme $V_1 \cup V_2$, $p = PJ(\pi_{\{V_1, V_2\}}(p))$ iff $V_1 - (V_1 \cap V_2)$ and $V_2 - (V_1 \cap V_2)$ are conditionally independent, given $V_1 \cap V_2$.

Let $K = \{p_1, \ldots, p_m\}$. The result of a sequence of applications

$$PJ\left(\ldots(PJ(PJ(p_{\sigma(1)} \cdot p_{\sigma(2)}), p_{\sigma(3)}), \ldots), p_{\sigma(m)}\right)$$

of the pairwise join procedure, where σ is a permutation of $\{1, \ldots, m\}$, is a *product extension* of K iff it is an element of $E(K)$. If p is a product extension of K, then $p = J(K)$ [19].

If $E(K)$ is nonempty and $s(K)$ is α-acyclic [10], then a (unique) product extension of K may be computed with σ corresponding to the reverse of any order in which elements of $s(K)$ are eliminated by Graham's algorithm [17]. Graham's algorithm [10] is a polynomial-time test for α-acyclicity of a model X.

1. $W := X$. Apply in any order until neither is applicable:
2. If v is an element appearing in only one $V \in W$, then $W := (W - \{V\}) \cup \{V - \{v\}\}$.
3. If $V_i \subseteq V_j$ for any $V_i, V_j \in W$, $i \neq j$, then $W := W - \{V_i\}$.

X is α-acyclic iff the algorithm terminates with $W = \{\emptyset\}$.

Example. $\{\{A, B\}, \{B, C\}\}$ is α-acyclic:

$$\{\{A, B\}, \{B, C\}\} \xrightarrow{2} \{\{B\}, \{B, C\}\} \xrightarrow{2} \{\{B\}, \{B\}\} \xrightarrow{3} \{\{B\}\} \xrightarrow{2} \{\emptyset\}.$$

$\{\{A, B\}, \{B, C\}, \{A, C\}\}$ is α-cyclic; neither step 2 nor step 3 is applicable.

For α-cyclic structures, an iterative proportional fitting algorithm (again, requiring only addition, multiplication and division) converges to $J(K)$ [7, 16].

A distribution p is *reconstructable* from a model X iff $p = J(\pi_X(p))$. If p is identifiable from X, then it is reconstructable from X, but not conversely.

In the context of reconstructability analysis [16], the problem of determining from a consistent database $\pi_X(p)$ as much as possible regarding the unknown distribution p is referred to as the *identification problem*. It is almost always the case that the system of projection equations (with unknowns $p(t)$) is underdetermined. So all that can be inferred deductively is that $p \in E(\pi_X(p))$. This may be sufficient for decision making (Section 4) or if bounds (determinable by linear programming) on particular $p(t)$ are all that is required.

The problem of identifying p from $\pi_X(p)$ is a type of *inverse problem*, in which data are generated via some non-injective mapping from a set of potential sources. The problem is to identify, using some reasonable criterion, a best representative element from the usually infinite set of preimages for the given data (in this case, a probabilistic database instance). In all published applications of reconstructability analysis, the solution has been to *maximize entropy* within $E(K)$; i.e., to select $J(E(K))$. The primary reason given is that this is the information-theoretically least bold inference that can be made from the data. Appeal is also made to Jaynes' *concentration theorem*, which has been interpreted as stating that the distribution $J(E(K))$ is the most likely to arise from observations satisfying the marginal constraints implied by K and that this likelihood decreases with increasing distance from $J(E(K))$ [13].

The maximum entropy approach is criticized in [21] and [32]. Interestingly, selection of the *centroid* of a set of distributions is advocated in [21]. The centroid, $C(E(K))$, minimizes the expected squared-error when it is selected as a solution to the identification problem. But $C(E(K))$ is more difficult to calculate than $J(E(K))$ [27]. When $X = \{V\}$ or $X = \{\emptyset\}$, $J(E(K)) = C(E(K))$; and in experiments involving approximately 8,000 randomly generated databases with non-trivial structures, the ratio of the squared-error distance between $J(E(K))$ and $C(E(K))$ to the squared-error diameter of $E(K)$ was found to be approximately 0.09 [28]. So, when selection of a single representative element of $E(K)$ is called for, $J(E(K))$ is not an unreasonable choice.

Perez [25] has explored the selection of the *barycenter* of a set of distributions S, the element of a set of distributions T minimizing the maximum distance d from it to any element of S. As a special case, when $Q = S = T$, a barycenter of a set Q of distributions is a $b \in Q$ such that

$$\min_{p \in Q} \max_{p' \in Q} d(p, p') = \max_{p' \in Q} d(b, p').$$

When Q is a convex polyhedron, for example, $E(Q)$ or $\pi_A(E(Q))$, and d is convex on Q,

$$\min_{p \in Q} \max_{p' \in Q} d(p, p') = \min_{p \in Q} \max_{p' \in L} d(p, p'),$$

where L is the set of vertices of Q [28]. Algorithms for determining such barycenters have been developed for various measures d by researchers in *location theory* [8].

3. ALGEBRAIC RESULTS

A number of results are presented now that are utilized in Section 4 to devise methods for decision-making with the information in a probabilistic database. Most of them are simple consequences of the definitions of Section 2.

Lemma 1. $\pi_V(p) = p$, if $V = s(p)$.

Lemma 2. $A \subseteq B$ implies $\pi_A(\pi_B(p)) = \pi_A(p)$.

Lemma 3. If $V = s(p)$ and $V \subseteq W \subseteq S$, then $\pi_W(E^S(p)) = E^W(p)$.

Lemma 4. If $s(K)$ is a cover of V, then $V \subseteq S$ implies $\pi_V(E^S(K)) \subseteq E(K)$.

Proof.

$$\begin{aligned}
\pi_V(E^S(K)) &= \pi_V\left(\bigcap_{p \in K} E^S(p)\right) && \text{[Def. Extension]} \\
&\subseteq \bigcap_{p \in K} \pi_V(E^S(p)) && \text{[Eq. (2)]} \\
&= \bigcap_{p \in K} E^V(p) && \text{[Lemma 3]} \\
&= E^V(K) && \text{[Def. Extension]}
\end{aligned}$$

□

Theorem 5. $X \leq Y$ implies $E^V(\pi_Y(p)) \subseteq E^V(\pi_X(p))$.

Proof [6]. $E^V(\pi_X(p))$ is the set of all solutions to the linear system determined by the projection of p onto the model X. If $X \leq Y$, then each equation determined by the projection of p onto X is a linear combination of equations in the system determined by the projection of p onto Y; thus, all solutions to the latter system are also solutions to the first. □

Theorem 5 has many useful consequences. For example, $X \leq Y$ implies $H(J(\pi_Y(p))) \leq H(J(\pi_X(p)))$, where H denotes entropy; thus, any refinement X of

a model Y of p embodies less information (quantified as negative entropy) about p than does Y. This is the basis for a number of model search procedures (in which the connectivity of a model reflects the dependency relations among the variables of the scheme for p) [9, 16].

Lemma 6. If $V_0 \subset \bigcup_{V \in s(K)} V$, then $\pi_{V_0}(p) \in \pi_{V_0}(E(K))$, for any $p \in E(K)$.

Where V_0 is the set of variables of actual concern (e. g., for a decision problem), it is not guaranteed, for arbitrary $p \in E(K)$, that $\pi_{V_0}(p)$ is contained in any sets smaller than $\pi_{V_0}(E(K))$ that can be constructed by means of the algebra, for example, $\pi_{V_0}(E(\pi_Y(p')))$, for $Y > s(K)$ and $p' \in E(K)$.

Theorem 7. Suppose $Y \leq X = s(K)$, and Y is a cover of V', where $V_0 \subseteq V' \subseteq V = \bigcup_{V_i \in X} V_i$. Then $\pi_{V_0}(E(K)) \subseteq \pi_{V_0}(E(\pi_Y(K)))$.

Proof. If $E(K) = \emptyset$, then $\pi_{V_0}(E(K)) = \emptyset$. If not, then $K = \pi_X(p)$ for any $p \in E(K)$, and

$$\begin{aligned}
\pi_{V_0}(E^V(\pi_X(p))) &= \pi_{V_0}(\pi_{V'}(E^V(\pi_X(p)))) && \text{[Lemma 2]} \\
&\subseteq \pi_{V_0}(\pi_{V'}(E^V(\pi_Y(p)))) && \text{[Theorem 5]} \\
&\subseteq \pi_{V_0}(E^{V'}(\pi_Y(p))) && \text{[Lemma 4]} \\
&= \pi_{V_0}(E^{V'}(\pi_Y(\pi_X(p)))) && \text{[Lemma 2]} \\
&= \pi_{V_0}(E(\pi_Y(K))). && \square
\end{aligned}$$

Corollary 8. For Z a cover of $V' \supseteq V_0$, and $Z \leq Y \leq s(K)$,

$$\pi_{V_0}(E(K)) \subseteq \pi_{V_0}(E(\pi_Y(K))) \subseteq \pi_{V_0}(E(\pi_Z(K))).$$

4. DECISION PROBLEMS

Arguably, the only reason to obtain or store probabilities is to base decisions on them. If probabilistic databases are to be useful (as relational databases are useful for problems of inventory, scheduling, etc.), it should be possible to devise algorithms for decision making from the information contained in them. The algebra of the previous section yields methods (of varying degrees of computational efficiency) supporting a variety of approaches to decision analysis.

Attention is restricted to decision problems consisting of a set $A = \{a_1, \ldots, a_m\}$ of mutually exclusive actions, a set $S = \{s_1, \ldots, s_n\}$ of mutually exclusive states, and a utility function $u : A \times S \longrightarrow R$. The set S will be constructible in some manner from the set $\mathrm{dom}(V_1 \cup \cdots \cup V_k)$ for a given probabilistic database $K = \{p_1, \ldots, p_k\}$.

Example. Let

$$A = \{\text{stay home, go swimming}\},$$

$S = \{$(rain, evening trains run), (rain, no trains), (no rain, trains), (no rain, no trains)$\}$. The utility function is given by the table

	stay home	go swimming
(rain, train)	3/4	1/2
(rain, no train)	7/8	0
(no rain, train)	1/8	1
(no rain, no train)	1/2	5/8

Probabilities are available in the form of a database including a variable for temperature and a variable indicating whether or not telephones are operating:

Rain	No_Phones	p_1	No_Phones	Temperature	p_2	Temperature	Trains	p_3
yes	true	3/8	true	high	1/3	high	yes	1/8
yes	false	1/16	true	med	5/36	high	no	3/8
no	true	1/8	true	low	1/36	med	yes	5/18
no	false	7/16	false	high	1/6	med	no	5/36
			false	med	5/18	low	yes	5/72
			false	low	1/18	low	no	1/72

Here,

$$\begin{aligned} S &= \mathrm{dom}\left(s\left(\pi_{\{\text{Rain, Trains}\}}(J(\{p_1, p_2, p_3\}))\right)\right) = \\ &= \mathrm{dom}\left(s\left(\pi_{\{\text{Rain, Trains}\}}(J(\{p_1, p_3\}))\right)\right) = \\ &= \mathrm{dom}\left(s\left(J\left(\{\pi_{\{\text{Rain}\}}(p_1),\ \pi_{\{\text{Trains}\}}(p_3)\}\right)\right)\right). \end{aligned}$$

But

$$\begin{aligned} \pi_{\{\text{Rain, Trains}\}}\left(J\left(\{p_1, p_2, p_3\}\right)\right) &\neq \pi_{\{\text{Rain, Trains}\}}\left(J\left(\{p_1, p_3\}\right)\right) \\ &\neq J\left(\{\pi_{\{\text{Rain}\}}(p_1),\ \pi_{\{\text{Trains}\}}(p_3)\}\right). \end{aligned}$$

Which of these probabilities should be used to calculate expected utilities? Should instead a *set* of probabilities be used? Which set?

There are several unproblematic cases in which probabilities p over S are obtained directly. If $S = \mathrm{dom}(V_i)$ for some $V_i \in \{V_1, \ldots, V_k\}$, then $p(s_j) = p_i(s_j)$. If $S = \mathrm{dom}(A)$ for $A \subseteq V_i$, $p(s_j) = \pi_A(p_i)(s_j)$. Similarly, if S is a partition of $\mathrm{dom}(V_i)$ or $\mathrm{dom}(A)$,

$$p(s_j) = \sum_{t \in s_j} p_i(t),$$

or

$$p(s_j) = \sum_{t \in s_j} \pi_A(p_i)(t).$$

In each case, the expected utility associated with an action a_i is calculated as

$$\sum_{j=1}^{n} p(s_j) \cdot u(a_i, s_j)$$

and an action maximizing expected utility is selected. (Of course, the sensitivity of the decision to slight variations in the originating p_i should be examined.)

In all other cases to be considered, $S = \mathrm{dom}(V_0)$, where $V_0 \subseteq V_1 \cup \cdots \cup V_k = V$ but there is no V_i such that $V_0 \subseteq V_i$. (Extension to related cases is straightforward.)

A distribution over $\mathrm{dom}(V_0)$ may always be computed from a consistent K, as $\pi_{V_0}(J(K))$. If $K = \pi_X(p)$ and it is known that $p = J(K)$, then this is the appropriate distribution. As discussed by Perez [24], this may be appropriate even when it is known that $p \simeq J(K)$, e.g., when

$$d(p, J(K)) \leq \varepsilon,$$

for some ε, where d is directed divergence (I-divergence, relative entropy):

$$d(p, q) = \sum_t p(t) \cdot \ln(p(t) / q(t)).$$

(Significant reductions in storage or transmission costs are possible if a small amount of error in reconstruction can be tolerated. In the extreme case, for a distribution p with a scheme V consisting of n k-valued attributes, representation of p requires k^n numbers vs. kn numbers for $\pi_X(p)$, where $X = \{\{v\} \mid v \in V\}$.)

Similarly, one may wish to compute the centroid or the barycenter of $E(K)$ and project onto V_0; the optimal action may be quite insensitive both to the choice of estimate and to perturbations of the estimate chosen. (Note that, in general, for $f_A : 2^{P_A} \longrightarrow P_A$, $\pi_{V_0}(f_V(E(K))) \neq f_{V_0}(\pi_{V_0}(E(K)))$.)

On the other hand, the structure of a database might not be determined by known relations of conditional independence which would allow confidence in the reconstruction (or approximate reconstruction) of a unique distribution. It might be impossible to obtain probabilities over the full set of variables of interest simultaneously, making it necessary to settle for a collection of distributions over various subsets. To give a fanciful example, imagine a study in which probabilities of occupation of the elements of a 3-dimensional grid are to be estimated for an object that can detect other objects only in the vertical ($\{X, Y\}$) plane. So as not to disrupt the normal pattern of movement, frequencies are recorded for the $\{Y, Z\}$ and $\{X, Z\}$ planes only. However, there is no reason to believe that location along the X coordinate is independent of location along Y, given the location along Z, and therefore no reason to believe that p over $\{X, Y, Z\}$ coincides with $J(\{\pi_{\{Y,Z\}}(p), \pi_{\{X,Z\}}(p)\})$.

Assuming correctness of the distributions $p_i \in K$, any element of the set $\pi_{V_0}(E(K))$ may be the actual distribution over V_0; i.e., the distribution relative to which, if it were known, one would choose an action with maximum expected utility. There are several methods of decision-making under partial uncertainty (i.e., when a probability distribution over S cannot be determined precisely) that may be adapted to this problem and that would not involve computing an estimate of this unknown distribution. That $\pi_{V_0}(E(K))$ is a convex polyhedron makes their application particularly feasible.

One method [18] retains for further consideration only actions for which there exists some $p \in \pi_{V_0}(E(K))$ relative to which it maximizes expected utility. Such actions are referred to as *E-admissible actions*, or *Bayes actions*. If there are multiple

Bayes actions, then non-probabilistic criteria (e. g., *maximin*) are applied to decide among them.

Let $K = \{p_1, \ldots, p_k\}$ with $V_0 \subset V = V_1 \cup \ldots \cup V_k$. An action a_i is E-admissible iff the linear system with

(i) $|\mathrm{dom}(V)| + |\mathrm{dom}(V_0)|$ unknowns: $p(x_1), \ldots, p(x_{|\mathrm{dom}(V)|})$ and $p(y_1), \ldots, p(y_{|\mathrm{dom}(V_0)|})$;

(ii) $\sum_{i=1}^{k} |\mathrm{dom}(V_i)|$ equations: $\sum_{x:\, x[V_i]=t} p(x) = p_i(t)$;

(iii) $|\mathrm{dom}(V_0)|$ equations: $\sum_{x:\, x[V_0]=y_h} p(x) - p(y_h) = 0$; and

(iv) $m - 1$ inequalities:

$$\sum_{h=1}^{|\mathrm{dom}(V_0)|} p(y_h) \cdot u(a_i, y_h) - \sum_{h=1}^{|\mathrm{dom}(V_0)|} p(y_h) \cdot u(a_j, y_h) \geq 0;$$

has a feasible solution.

For any $p \in \pi_{V_0}(E(K))$, let

$$e_p(a) = \sum_y p(y) \cdot u(a, y).$$

Since $\pi_{V_0}(E(K))$ is convex, the set

$$U(a) = \{e_p(a) \mid p \in \pi_{V_0}(E(K))\}$$

is an interval. Its endpoints are calculated from the linear system consisting of (i), (ii) and (iii), above, and the objective function $e_p(a)$.

A less stringent elimination criterion than E-admissibility is based on the ordering

$$a_i > a_j \quad \text{iff } \min U(a_i) > \max U(a_j).$$

Only the maximal elements of A under this ordering (i. e., actions a_i such that there is no a_j for which $a_j > a_i$) are retained as admissible [20]. (If a is E-admissible, then it is maximal under '$>$', but not conversely.)

More radical criteria utilizing intervals $U(a)$ include a *generalized* (or *gamma-*) *maximin criterion* [11],

$$\text{choose } \arg\max_{a \in A} \min U(a);$$

and *generalized Hurwicz criterion* [12],

$$\text{choose } \arg\max_{a \in A} \alpha \min U(a) + (1 - \alpha) \max U(a), \quad 0 \leq \alpha \leq 1.$$

The linear system above is also the basis for application to such problems of the *domain criterion* [31]. The domain of $a \in A$ is the set

$$D(a) = \{p \in \pi_{V_0}(E(K)) \mid e_p(a) \geq e_p(a') \text{ for all } a' \in A\}.$$

(Note that $D(a) = \emptyset$ iff a is not E-admissible.) Assuming uniform probability of "correctness" over $E(K)$, the ratio of the volume of $D(a)$ to the volume of $\pi_{V_0}(E(K))$ may be interpreted as the probability that action a maximizes expected utility relative to the actual but unknown distribution $\pi_{V_0}(p)$. Selecting the action with largest domain is then selection of the action that is likeliest to maximize expected utility.

Application of any of these criteria to the full system of linear inequalities may be needlessly expensive. It may be possible to obtain the same result with a refinement $\pi_Y(K)$ of the initially given database K for which the size of the components (i) and (ii) of the linear system is reduced. Consider the E-admissibility criterion with fixed V_0, K, A, and u. Suppose there exists a uniquely E-admissible action relative to the information in K. By Theorem 7, if there exists a uniquely E-admissible action with the substitution of $\pi_Y(K)$ for K, where Y is a cover of $V' \supseteq V_0$, then the two actions coincide. (Similarly for the utility interval dominance criterion.)

This suggests the following strategy: Starting with $W = \{\{v\} \mid v \in V_0\}$, the most refined model that is a cover of some $V' \supseteq V_0$, repeatedly aggregate W until there is a single admissible action relative to $\pi_{V_0}(E(\pi_W(K)))$, or $\pi_{V_0}(E(\pi_W(K)))$ happens to contain only one distribution, or $W = s(K)$, whichever comes first. However, if models are replaced by immediate aggregation, very little progress toward sufficient reduction of the set $\pi_{V_0}(E(\pi_W(K)))$ is likely to be made at each step. Also, there will not be a unique immediate aggregate at each step. A reasonable alternative is the sequence of models [29]:

$$(\{\{v\} \mid v \in V_0\}, \{V_i \cap V_0 \mid V_i \in s(K),\ V_i \cap V_0 \neq \emptyset\}, \{V_i \mid V_i \in s(K),\ V_i \cap V_0 \neq \emptyset\}, s(K)).$$

Following this strategy, suppose that a single admissible action is identified with $W = \{\{v\} \mid v \in V_0\}$. Then linear programs with only $|\mathrm{dom}(V_0)|$ unknowns and $m - 1$ inequalities are sufficient. (Compare with (i) – (iii) above.) The set $|\mathrm{dom}(V)|$ can be arbitrarily large, and many (or all) of the variables in $V - V_0$ may be irrelevant to the given problem.

If the conjecture below is correct, then there exists a polynomial-time algorithm for eliminating irrelevant attributes from a connected α-acyclic database.

For $s(K)$ a cover of $V \supseteq V_0$, define a bipartite graph B with node set $V \cup s(K)$ and edge set

$$\{\{v, x\} \mid v \in V,\ x \in s(K),\ v \in x\}.$$

Let

$$R = \{v \in V \mid v \in V_0 \text{ or } v \text{ is on an acyclic path in } B \text{ between two elements of } V_0\}.$$

Let

$$Z = \{V_i \cap R \mid V_i \in s(K) \text{ and } V_i \cap R \not\subseteq V_j \cap R, \text{ for all } V_j \in s(K),\ i \neq j\}.$$

Conjecture: $\pi_{V_0}(E(\pi_Z(K))) = \pi_{V_0}(E(K))$.

If $s(K)$ is connected and α-acyclic, then the following polynomial-time algorithm will produce the reduced structure Z [23]:

1. $Z := s(K)$.
2. Repeat in any order until neither has any effect on the current value of Z:
   a. If a variable $v \notin V_0$ appears in only one element of Z, remove v from that element.
   b. If Z contains elements V_i and V_j such that $V_i \subset V_j$, then $Z := Z - \{V_i\}$.

For the E-admissibility and utility interval dominance criteria, the truth of the conjecture would permit substitution of Z for $s(K)$ in the sequence of aggregates above. Or Z may be refined enough that it is reasonable to work directly with $\pi_{V_0}(E(\pi_Z(K)))$. Methods in which expected utility is to be maximized relative to an element $f_{V_0}(\pi_{V_0}(E(K)))$ would also gain in efficiency if $\pi_{V_0}(E(K)) = \pi_{V_0}(E(\pi_Z(K)))$ by reduction of the number of unknowns in the linear system characterizing the set from which the element is to be selected.

5. UPDATING PROBABILISTIC DATABASES

Probabilistic databases may be updated when new information is received.

Let us reconsider the decision problem utilizing the database

Rain	No_Phones	p_1	No_Phones	Temperature	p_2	Temperature	Trains	p_3
yes	true	3/8	true	high	1/3	high	yes	1/8
yes	false	1/16	true	med	5/36	high	no	3/8
no	true	1/8	true	low	1/36	med	yes	5/18
no	false	7/16	false	high	1/6	med	no	5/36
			false	med	5/18	low	yes	5/72
			false	low	1/18	low	no	1/72

We may imagine that the database contains probabilities for a "typical" August day. Suppose that we measure the temperature on the particular August afternoon during which we are trying to decide whether or not to go swimming and find that it is in the low range; i.e., we are certain that the value of the variable Temperature is "low". We may update by ordinary conditionalization our probabilities for the variables No_Phones and Trains; for example:

$$p_2'(\text{No\_Phones} = \text{true}) = p_2(\text{No\_Phones} = \text{true} \mid \text{Temperature} = \text{low})$$
$$= \frac{1/36}{1/36 + 1/18} = 1/3.$$

Now the probabilities in the original distribution p_1 must be reconciled with the new probabilities for No_Phones. This may be done by means of *Jeffrey's Rule* [15]: If conditional probabilities $p(A|B_i)$ are unaffected by a change in probabilities for

mutually exclusive and exhaustive events $B_1, \ldots, B_n$, then, in light of these changes, the probability of event A should be updated to

$$p'(A) = \sum_i p'(A|B_i) \cdot p'(B_i) = \sum_i p(A|B_i) \cdot p'(B_i).$$

We may assume that conditional probabilities

$$p(\text{Rain} = x \mid \text{No\_Phones} = z)$$

are unaffected by the change in probabilities for the propositions "No_Phones = yes" and "No_Phones = no", which in turn was necessitated by our observing "Temperature = low". Our database, from which these and many other conditional probabilities may be calculated, allows for the possibility that the temperature on an August day may be low. (Our database for a typical December day, however, may embody very different conditional probabilities; low temperatures in December may make telephone problems *more* likely than when temperatures are in the medium or high ranges.) So, referring to an arbitrary $p \in E(\{p_1, p_2, p_2\})$ and $p' \in E(\{p'_1, p'_2, p'_3\})$:

$$\begin{aligned} p'(\text{Rain} = x,\ \text{No\_Phones} = z) &= p'(\text{Rain} = x \mid \text{No\_Phones} = z) \cdot p'(\text{No\_Phones} = z) \\ &= p(\text{Rain} = x \mid \text{No\_Phones} = z) \cdot p'(\text{No\_Phones} = z) \end{aligned}$$

Since Temperature should no longer be regarded as a variable, the updated database is:

Rain	No_Phones	p'_1		No_Phones	p'_2	Trains	p'_3
yes	true	1/4	$= 3/4 \cdot 1/3$	true	1/3	yes	5/6
yes	false	1/12	$= 1/8 \cdot 2/3$	false	2/3	no	1/6
no	true	1/12	$= 1/4 \cdot 1/3$				
no	false	7/12	$= 7/8 \cdot 2/3$				

(Note that the structure of this database is not a model.)

Let us now apply the conjecture stated in Section 4.

With $s(K) = \{\{\text{Rain, No\_Phones}\}, \{\text{No\_Phones}\}, \{\text{Trains}\}\}$ and $V_0 = \{\text{Rain, Trains}\}$, Z is the set $\{\{\text{Rain}\}, \{\text{Trains}\}\}$. Projecting onto Z yields the database

Rain	$\pi_{\{\text{Rain}\}}(p'_1)$	Trains	p'_3
true	1/3	yes	5/6
false	2/3	no	1/6

We do *not* assume probabilistic independence of the variables Rain and Trains. Let us apply the E-admissibility criterion relative to the set of probability distributions $E(\{\pi_{\{\text{Rain}\}}(p'_1), p'_3\})$. There exist solutions to the system of inequalities

$$\begin{aligned} &p(\text{Rain} = \text{yes, Trains} = \text{yes}) + p(\text{Rain} = \text{yes, Trains} = \text{no}) = 1/3 \\ &p(\text{Rain} = \text{no, Trains} = \text{yes}) + p(\text{Rain} = \text{no, Trains} = \text{no}) = 2/3 \\ &p(\text{Rain} = \text{yes, Trains} = \text{yes}) + p(\text{Rain} = \text{no, Trains} = \text{yes}) = 5/6 \\ &p(\text{Rain} = \text{yes, Trains} = \text{no}) + p(\text{Rain} = \text{no, Trains} = \text{no}) = 1/6 \\ &p(\text{Rain} = \text{yes, Trains} = \text{yes} \cdot (1/2 - 3/4) \\ &\quad + p(\text{Rain} = \text{yes, Trains} = \text{no} \cdot (0 - 7/8) \\ &\quad + p(\text{Rain} = \text{no, Trains} = \text{yes} \cdot (1 - 1/8) \\ &\quad + p(\text{Rain} = \text{no, Trains} = \text{no} \cdot (5/8 - 1/2) \geq 0. \end{aligned}$$

Thus, there exist probability distributions in $E(\{\pi_{\{\text{Rain}\}}(p_1'), p_3'\})$ relative to which the action "go swimming" maximizes expected utility. The corresponding system of inequalities for the action "stay home" has no feasible solutions. Therefore, "go swimming" is uniquely E-admissible.

6. CONCLUSION AND OPEN PROBLEMS

Clearly, determining the truth or falsity of the conjecture of Section 4 is an important problem. If false, an alternative method of extracting from a database information relevant to a given decision problem is needed. If it is true, then an efficient method for constructing the reduced structure must be sought when the original database structure is not connected and α-acyclic.

A fundamental issue is whether working with the entire set of distributions of which the elements of a database are marginals or with a single "best representative" (barycenter, centroid, maximum entropy element) of this set is the more appropriate for decision making under the form of uncertainty represented by a probabilistic database.

If the latter course is taken, then, in light of the failure of the marginalization property

$$\pi_{V_0}(f_V(E(K)) = f_{V_0}(\pi_{V_0}(E(K)))$$

to hold in general, a secondary issue that requires investigation is the appropriateness of solving a given decision problem over a set of variables V_0 relative to one vs. any other of these estimates of the unknown distribution.

The probabilistic data model supports each of these approaches. As discussed in Section 4, the linear inequalities from which it can be determined whether or not an action is a Bayes solution (is E-admissible) relative to a database are easily set up. At the same time, determining a representative joint distribution of a set of distributions (in particular, the maximum entropy distribution) is more efficient when the constraints defining the set are in the form of a probabilistic database than in the general case.

ACKNOWLEDGEMENT

The author is grateful to Dr. Otakar Kříž for many helpful suggestions.

(Received November 28, 1991.)

REFERENCES

[1] J. Aczel and Z. Daroczy: On Measures of Information and their Characterizations. Academic Press, New York 1975.

[2] D. Barbara, H. Garcia-Molina and D. Porter: The management of probabilistic data. IEEE Trans. Knowl. Data Engrg., to appear.

[3] C. Berge: Graphs and Hypergraphs. North Holland, Amsterdam 1973.

[4] R. Cavallo: Lattices of structure models and database schemes. Internat. J. Gen. Systems *20* (1992), 247–274.

[5] R. Cavallo and G. Klir: Reconstructability analysis of multi-dimensional relations: a theoretical basis for computer-aided determination of acceptable systems models. Internat. J. Gen. Systems *5* (1979), 143–171.
[6] R. Cavallo and M. Pittarelli: The theory of probabilistic databases. Proc. 13th Internat. Conf. Very Large Databases, 1987, pp. 71–81.
[7] W. Deming and F. Stephan: On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. *11* (1940), 427–444.
[8] W. Domschke and A. Drexl: Location and Layout Planning: an international bibliography. Springer – Verlag, New York 1985.
[9] D. Edwards and T. Havránek: A fast procedure for model search in multidimensional contingency tables. Biometrika *72* (1985), 339–351.
[10] R. Fagin: Degrees of acyclicity for hypergraphs and relational database schemes. J. Assoc. Comput. Mach. *30* (1983), 514–550.
[11] P. Gardenfors: Forecasts, decisions and uncertain probabilities. Erkenntnis *14* (1979), 159–181.
[12] L. Hurwicz: Problems of incorrect and incomplete specification. Econometrica *19* (1951), 343–344.
[13] E. T. Jaynes: On the rationale of maximum-entropy methods. Proc. IEEE *70* (1982), 939–952.
[14] E. T. Jaynes: Prior information and ambiguity in inverse problems. SIAM-AMS Proceedings *14* (1984), 151–166.
[15] R. C. Jeffrey: The Logic of Decision. Second edition. University of Chicago Press, Chicago 1983.
[16] G. Klir: Architecture of Systems Problem Solving. Plenum Press, New York 1985.
[17] K. Krippendorff: Information Theory: structural models for qualitative data. Sage Publications, Beverly Hills 1986.
[18] I. Levi: The Enterprise of Knowledge. MIT Press, Cambridge, Mass. 1980.
[19] P. Lewis: Approximating probability distributions to reduce storage requirements. Inform. and Control *2* (1959), 214–225.
[20] R. Loui: Decisions with indeterminate probabilities. Theory and Decision *21* (1986), 283–309.
[21] J. MacQueen and J. Marschak: Partial knowledge, entropy, and estimation. Proc. Nat. Acad. Sci. *72* (1975), 3819–3824.
[22] D. Maier: The Theory of Relational Databases. Computer Science Press, Rockville, MD 1983.
[23] D. Maier and J. Ullman: Connections in acyclic hypergraphs. Proc. ACM Symp. on Principles of Database Systems, 1982, pp. 34–39.
[24] A. Perez: ε-admissible simplifications of the dependence structure of a set of random variables. Kybernetika *13* (1977), 439–449.
[25] A. Perez: The barycenter concept of a set of probability measures as a tool in statistical decision. In: Probability Theory and Mathematical Statistics, vol. 2, VNU Science Press 1986, pp. 437–450.
[26] A. Perez and R. Jiroušek: Constructing an intensional expert system (INES). In: Medical Decision Making: diagnostic strategies and expert systems, North Holland, Amsterdam 1985, pp. 307–315.
[27] F. Piepel: Calculating centroids in constrained mixture experiments. Technometrics *25* (1983), 279–283.
[28] M. Pittarelli: Uncertainty and estimation in reconstructability analysis. Internat. J. Gen. Systems *18* (1989), 1–58.
[29] M. Pittarelli: An algebra for probabilistic databases. IEEE Trans. Knowl. Data Engrg. To appear 1993.

[30] K. Raju and A. Majumdar: The study of joins in fuzzy relational databases. Fuzzy Sets and Systems *21* (1987), 19–34.
[31] G. Schneller and G. Sphicas: Decision making under uncertainty: Starr's domain criterion. Theory and Decision *15* (1983), 321–336.
[32] T. Seidenfeld: Entropy and uncertainty. Philosophy of Science *53* (1986), 467–491.

Dr. Michael Pittarelli, Computer Science Department, State University of New York Institute of Technology, Utica, NY 13504-3050. U. S. A.