ISSN 0980-1472

Revue Internationale de

systēmique

Vol. 4, N° 1, 1990

afcet

Dunod

AFSCET

Revue Internationale de systēmique

Revue Internationale de Sytémique

volume 04, numéro 1, page 45 - 64, 1990

Data Synthesis from Probabilistic Structure Systems

Michael Pittarelli

Numérisation Afscet, janvier 2016.

REVUE INTERNATIONALE DE SYSTEMIQUE
Vol. 4, N° 1, 1990, pp. 45 à 64

DATA SYNTHESIS FROM PROBABILISTIC STRUCTURE SYSTEMS

Michael PITTARELLI
State University of New York Institute of Technology [1]

Abstract
Methods of constructing data from the information in a probabilistic structure system are presented. The resulting data may take various forms: a single probability distribution, bounds on components of a single distribution, a set of distributions, or another structure system. Rudiments of an algebra for probability distributions over finite structured spaces are developed by means of which the correctness of several data synthesis procedures is demonstrated. The potential usefulness of these procedures is illustrated through their application to problems of decision making under uncertainty.

Résumé
Nous présentons des techniques pour construire des données à partir d'un système de structures aléatoires. Les données qui en résultent peuvent prendre des formes diverses : loi de probabilité unique, bornes concernant les composantes d'une loi de probabilité unique, ensemble de lois de probabilité, ou un autre système de structures. Nous développons les rudiments d'une algèbre pour les lois de probabilité définies sur des espaces à structure finie, ce qui nous permet de démontrer la validité de quelques méthodes de synthèses de données. Nous faisons valoir l'utilité possible de ces méthodes en les appliquant à quelques problèmes de choix de comportement rationnel face au risque.

1. Introduction: structure systems

A probabilistic structure system is, essentially, a collection of probability distributions over Cartesian products of finite variable domains. Formally, a

1. Utica, New York 13504-3050, U.S.A., Department of Computer and Information Science.

Rev. intern. systémique. 0980-1472 Vol. 4/90/01/ 45 /20/$ 4.00/

structure system may be defined as a set $S=\{P_1, \ldots, P_m\}$ of *probabilistic systems*. A *probabilistic system* is a four-tuple $P=(V, \Delta, \text{dom}, p)$, where

- V is a non-empty set of *variables* (or *attributes*), and V is referred to as the *scheme* for P;
- Δ is a non-empty set of finite sets of values called *domains;*
- dom: $V \to \Delta$ is an onto function that associates a domain with each variable;
- $\text{dom}(A) = \underset{v \in A}{\times} \text{dom}(v)$, where $A \subseteq V$, is the set of system substates (subtuples) over variables A [dom (V) is the set of system states, or system tuples];
- for $w \in \text{dom}(W)$, $b \in \text{dom}(B)$, and $B \subseteq W$, $w[B]=b$ iff b and w agree on all attributes in B;
- p : dom (V) $\to [0, 1]$, where $\sum_{t \in \text{dom}(V)} p(t)=1$ is a probability distribution over V. [For convenience, a total ordering on dom (V) is assumed, and p is used to denote the correspondingly ordered tuple of images of the function p. Also, since p determines V, Δ, and dom, "p" is sometimes used to refer to the system P.]

Example 1.1. – The tables below represent a structure system $S=\{P_1, P_2, P_3\}$ for which $V_1=\{A, B\}$, $\Delta_1=\{\{a_1, a_2\}, \{b_1, b_2\}\}$, dom $(A)=\{a_1, a_2\}$, $\text{dom}(B)=\{b_1, b_2\}$, $V_2=\{B, C\}$, etc.

A	B	$p_1(.)$	B	C	$p_2(.)$	D	$0_3(.)$
a_1	b_1	0	b_1	c_1	0.1	d_1	0.25
a_1	b_2	0.2	b_1	c_2	0.3	d_2	0.25
a_2	b_1	0.4	b_2	c_1	0.2	d_3	0.5
a_2	b_2	0.4	b_2	c_2	0.4		

A structure system $\{P_1, \ldots, P_m\}$ has the *structure* $X=\{V_1, \ldots, V_m\}$, where V_i is the scheme for P_i. For the example above, the structure is $\{\{A, B\}, \{B, C\}, \{D\}\}$.

A probabilistic structure system may be viewed as a *probabilistic database*. Mathematically, all that distinguishes a probabilistic structure system from a relational database instance is the nature of the functions p_i. If they were to be replaced by characteristic functions r_i : dom (V) $\to \{0, 1\}$, the result would be a relational database. (In the reconstructability analysis literature, what is referred to as a "set-theoretic" structure system is identical to a relational database instance. The term "structure system" will be used in the remainder of the paper to refer to probabilistic structure systems.) A previous paper (Cavallo and Pittarelli, 1987) discusses probabilistic analogues of relational data dependencies and information-preserving transformations of relational

to probabilistic database instances. (The paper also mentions the priority of work in the areas of dependency theory, decomposition theory for relations, etc., performed by systems theorists, most notably W. R. Ashby. In the computer science literature, the origins of such investigations are erroneously attributed to E. F. Codd.)

The purpose of this paper is to describe methods of synthesizing data from the information contained in a probabilistic structure system. The resulting data may take various forms: a single probability distribution, a set of probability distributions, an "interval-valued" distribution, or another structure system. The process of synthesizing data in this sense is similar to the construction of "views" for relational databases. Just as procedures for view construction require for their precise definition (as well as for optimization) an algebra (relational algebra) for manipulating relational data, so too is an algebra for probabilistic data necessary.

2. Probabilistic algebra

The basic operations required for the procedures to be discussed are defined in this section. (None of them is new. They are redefined here for completeness.) Algebriac facts relating the various operations are then presented, mostly without proof, as a series of propositions. These will be used in demonstrations of the correctness of various data synthesis procedures.

Definition 2.1. – The *projection* of a distribution p with scheme V onto $A \subseteq V$ is the distribution $\pi_A(p)$, where

$$\pi_A(p) : \text{dom}(A) \to [0,1]$$

and

$$\pi_A(p)(a) = \sum_{s[A]=a} p(s). \quad \square$$

The projection of a distribution onto a subset of its scheme is unique. (In the probability and statistics literature, the term "marginalization" is used instead of "projection".)

Example 2.1. – Projecting distribution p_2 of the structure system in Example 1.1 onto the singleton set $\{C\}$ of variables yields the distribution

C	$\pi_{\{C\}}(p_2)(.)$
c_1	0.3
c_2	0.7

Definition 2.2. – A *model* of a set of variables V is a structure $X=\{V_1, \ldots, V_m\}$ such that $\bigcup_{j=1}^{m} V_j \subseteq V$ and $V_i \not\subseteq V_j$ for all $i, j \in \{1, \ldots, m\}$. (X needn't be a cover of V.)

Projecting p with scheme V onto a model $X=\{V_1, \ldots, V_m\}$ of V results in the structure system

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

Example 2.2. – The subset $S'=\{P_1, P_3\}$ of the structure system $S=\{P_1, P_2, P_3\}$ in Example 1.1 is the projection onto the structure $\{\{A, B\}, \{D\}\}$ of the distribution p below:

A	B	D	$p(.)$
a_1	b_1	d_1	0
a_1	b_1	d_2	0
a_1	b_1	d_3	0
a_1	b_2	d_1	0.05
a_1	b_2	d_2	0.05
a_1	b_2	d_3	0.1
a_2	b_1	d_1	0.1
a_2	b_1	d_2	0.1
a_2	b_1	d_3	0.2
a_2	b_2	d_1	0.1
a_2	b_2	d_2	0.1
a_2	b_2	d_3	0.2

The *refinement relation* (Cavallo and Klir, 1979) between structures is important not only for data analysis (e.g., reconstructability analysis) but also for data synthesis.

Definition 2.3. – A structure X is a *refinement* of structure Y, denoted $X \subseteq Y$, iff for each $V_x \in X$ there exists a $V_y \in Y$ such that $V_x \subseteq V_y$. For example, $\{\{A\}, \{B, C\}\}$ is a refinement of $\{\{A, B\}, \{B, C\}, \{D\}\}$. $\square$

A structure system S with structure Y may be projected onto a refinement X of Y to form a structure system $\pi_X(S)$ each element of which is a projection of some element of S.

Example 2.3. – The projection of $S=\{P_1, P_2, P_3\}$ of Example 1.1 onto the structure $\{\{A\}, \{B, C\}\}$ results in the structure system $\pi_{\{\{A\}, \{B, C\}\}}(S)$

with elements

A	$\pi_{\{A\}}(p_1)(.)$	B	C	$p_2(.)$
a_1	0.2	b_1	c_1	0.1
a_2	0.8	b_1	c_2	0.3
		b_2	c_1	0.2
		b_2	c_2	0.4

Note that the second element of the structure system above has as its distribution the original distribution p_2 of the operand structure system. This is a consequence of

Proposition 1. – $\pi_V(p)=p$, *if* V *is the scheme for* p. $\square$

As noted previously, a structure system S may be constructed by projection of a known probability distribution p with scheme V onto a model X of V. If such a structure system is itself projected onto a refinement Y of X, the resulting system $\pi_Y(S)$ is identical to the system $\pi_Y(p)$ that would have resulted from projecting p directly onto Y.

Theorem 1. – $Y \subseteq X$ *implies* $\pi_Y(\pi_X(p))=\pi_Y(p)$.

Proof. – Follows directly from:

Proposition 2. – $A \subseteq B$ *implies* $\pi_A(\pi_B(p))=\pi_A(p)$. $\square$

The operator symbol "π" is *overloaded* in the sense that it represents operators of different types, where the operator type depends on the types of the two operands in expressions $\pi_{o_1}(o_2)$ For example, when o_1 is a structure and o_2 is a distribution, the domain of π is the Cartesian product of a set of structures with a set of distributions, and the codomain is a set of structure systems. It will frequently be convient, as suggested by the notation, to view π_{o_1} as the name of a unary function of o_2 in expressions $\pi_{o_1}(o_2)$. Thus, $\pi_{o_1}(A_2)$, where A_2 is a *set* of distributions or structure systems, denotes the image of A_2 under the mapping π_{o_1} (Bourbaki, 1954):

$$\pi_{o_1}(A_2)=\{\pi_{o_1}(o_2) \mid o_2 \in A_2\}.$$

In what follows there will be occasional use of projection expressions of this last type.

For a given system S with structure $X=\{V_1, \ldots, V_m\}$ there is, in general, an infinite set of distributions p over any scheme V for which X is a model such that $\pi_X(p)=S$. Thus, mappings π_X are usually not invertible. This is the crux of the *identification problem* of reconstructability analysis: given a structure system S, characterize in some way the set of distributions p whose

projection onto the structure of S yields S (where a particular scheme V for p, usually the union of the elements of the structure of S, is assumed).

Definition 2.4. – For a system $S=\{p_1, \ldots, p_m\}$ whose structure $X=\{V_1, \ldots, V_m\}$ is a model of V, the *extension* of S over V is the set

$$E_V(S)=\{p\in P^n \mid \pi_{V_i}(p)=p_i,\ i=1, \ldots, m\},$$

where $n=|\mathrm{dom}(V)|$ and P^n is the simplex of all n-component probability distributions on dom(V). If $V=V_1\cup\ldots\cup V_m$, then $E_V(S)$ may be written E(S), and coincides with the standard reconstruction family of S. □

Example 2.4. – The extension of the structure system $\{p_1, p_2\}$ below

A	$p_1(.)$	C	$p_2(.)$
a_1	0.2	c_1	0.3
a_2	0.8	c_2	0.7

over the scheme $\{A, C\}$ is the set $E(\{p_1, p_2\})$ of solutions p to the system of equations and inequalities

$$p(a_1 c_1)+p(a_1 c_2)=0.2$$
$$p(a_2 c_1)+p(a_2 c_2)=0.8$$
$$p(a_1 c_1)+p(a_2 c_1)=0.3$$
$$p(a_1 c_2)+p(a_2 c_2)=0.7$$
$$p(.)\geqq 0.$$

(Note that the equations imply that $p(a_1 c_1)+\ldots+p(a_2 c_2)=1$.)

Extension is complementary to projection in the weak sense that:

Proposition 3. – $\pi_V(p)\in E_V(\pi_X(p))$.

When V is the scheme for p this reduces, by Proposition 1, to $p\in E(\pi_X(p))$. □

Definition 2.5. – p over V is *identifiable from* X iff $E_V(\pi_X(p))=\{p\}$. □

Some additional algebraic facts relating the operations of projection and extension follow.

Proposition 4. – $E_V(\pi_{\{\varnothing\}}(p))=P^n$. □

Proposition 5. – $E_V(\pi_{\{V\}}(p))=\{\pi_V(p)\}$. □

Proposition 6. – $E_V(\pi_X(p))$ *is a proper subset of* P^n, *for any* $X\neq\{\varnothing\}$. □

Proposition 7. – $V\subseteqq Q$ *implies* $\pi_V(E_Q(S))=E_V(S)$, *where* $\pi_V(E_Q(S))$ *denotes the range of the function* π_V *applied to the elements of* $E_Q(S)$, *and* S's *structure is a cover of* V. □

Proposition 8. – $\pi_V(E_V(S))=E_V(S)$. □

Proposition 9. – $\pi_X(E(\pi_X(p)))=\{\pi_X(p)\}$. □

Proposition 10. – $X\subseteqq Y$ *implies* $E_V(\pi_Y(p))\subseteqq E_V(\pi_X(p))$.

Proposition 10 is easily proved (Cavallo and Pittarelli, 1987) by noting that any $\pi_Z(p)$, where Z is a scheme, represents a set of linear equations and inequalities. $E_V(\pi_X(p))$ represents the set of all solutions to the linear system determined by the projection of p onto elements of the structure X. If $X\subseteqq Y$, then each equation determined by the projection of p onto elements of X is a linear combination of equations in the system determined by the projection of p onto elements of Y; thus, all solutions to the latter system are also solutions to the first; *i.e.*, $E_V(\pi_Y(p))\subseteqq E_V(\pi_X(p))$. □

The maximum and minimum values, as p ranges over $E_V(S)$, of $p(t)$ for a given $t\in\mathrm{dom}(V)$ can be determined by linear programming. These values are the endpoints of *probability intervals* consisting of all possible values for $p(t)$, given S. For the structure system $\{p_1, p_2\}$ of Example 2.4, these intervals, $i(t)$, are

A	C	$i(.)$
a_1	c_1	[0,0.2]
a_1	c_2	[0,0.2]
a_2	c_1	[0.1,0.2]
a_2	c_2	[0.5,0.7]

The vector of intervals i may be viewed as an "interval-valued" probability distribution. For an individual tuple t considered in isolation from all other $t'\in\mathrm{dom}(V)$, the interval $i(t)$ contains all the information regarding its probability that can be deduced from S. Real-valued (*i.e.*, ordinary) distributions p can be *inductively* inferred from S but, unless this p is *identifiable* from its projections onto the scheme for S, *i.e.*, $E(S)=\{p\}$, the actual joint distribution over the variables V may not be the distribution p inferred. As will be illustrated, and as is discussed more fully elsewhere (Seidenfeld, 1983; Pittarelli, 1988, 1989), interval valued distributions and the sets of real-valued distributions $E_V(S)$ can be put to many of the same uses as single real-valued distributions. Thus, it is not always necessary to estimate a single $p\in E_V(S)$. The narrower the components of an interval-valued distribution or the smaller a set of distributions $E_V(S)$, the closer the results of procedures (e.g., decision procedures) devised for use with them will be to those of procedures suitable

for single real-valued distributions. The goal of many of the data synthesis techniques to be discussed is the minimization of these widths and sizes.

A *join* operation has as domain a set of structure systems and as codomain a set of probability distributions. (The term derives from relational database theory, where a join operation applied to a relational database instance yields a "universal" relation instance compatible with it.)

The result of applying a join operator J to a structure system S is a distribution such that its projection onto the structure of S yields S. Formally, where X is the structure of S,

$$\pi_X(J(S)) = S.$$

Equivalently, for any p and any cover X of its scheme,

$$\pi_X(J(\pi_X(p))) = \pi_X(p).$$

In other words,

$$J(\pi_X(p)) \in E(\pi_X(p)).$$

The composite function $J \circ \pi_X$, the *project-join* operation, maps distributions to distributions.

Definition 2.6. – A distribution is *reconstructable*, relative to a project-join operator $J \circ \pi_X$, iff it is a *fixed point* of $J \circ \pi_X$, *i.e.*, iff

$$J(\pi_X(p)) = p. \quad \square$$

(Observe that identifiability of p from X implies reconstructability from X, for any join operator J, but that reconstructability does not imply identifiability.) For the join operations to be discussed, there is a unique element p of any equivalence class (with respect to projection onto X) $E(\pi_X(p))$ such that $p = J(\pi_X(p))$. Thus, the property of (perfect vs. approximate) reconstructability is extremely rare.

A widely used measure of the uncertainty represented by a finite probability distribution is (Shannon) entropy,

$$H(p) = -\sum_t p(t) \log_2 p(t).$$

For any reconstruction family E(S), there is a unique maximum entropy element (Cavallo and Klir, 1981; Jaynes, 1984). The canonical *probabilistic join* procedure of reconstructability analysis yields the (unique) maximum entropy element, p^*, of a reconstruction family (Cavallo and Klir, 1981): $J(E(S)) = p^*$. The *maximum entropy principle* of inductive inference prescribes estimation of $p(t)$ as $p^*(t)$ when one has information regarding $t \in \mathrm{dom}(V)$ in the form of a structure system S over a cover of V. (This principle is discussed critically in [Pittarelli, 1989].)

Alternatives to the maximum entropy join have been considered (Pittarelli, 1989): the *centroid* of E(S), or a minimax distance join, $J(E(S)) = p_m$, where

$$\max_{p' \in E(S)} d(p_m, p') = \min_{p \in E(S)} \max_{p' \in E(S)} d(p, p'),$$

for a suitable distance measure d. However, only the maximum entropy join has been employed in practice. Thus, in what follows, the term "join" will refer to the maximum entropy join operation.

3. Data synthesis

Suppose that a distribution p with scheme V is reconstructable from some structure $X = \{V_1, \ldots, V_m\} : p = J(\pi_X(p))$. Further, suppose that the structure system $S = \pi_X(p)$ is stored, and not p itself. When X is sufficiently refined, the storage and transmission cost savings achieved thereby can be dramatic. For example, if $|V| = n$ and $|\mathrm{dom}(v)| = k$ for all $v \in V$, explicit storage of p involves k^n numbers. Storage of $\pi_X(p)$, for $X = \{\{v\} \mid v \in V\}$, requires only kn numbers. [This was one of the motivations behind early work in this area (Lewis, 1959).]

Assume now that the projection of p onto the structure $Y = \{V_i, \ldots, V_j\}$ is desired, where $V_i \cup \ldots \cup V_j \subseteq V_1 \cup \ldots \cup V_m$. One obvious strategy, since p is reconstructable from S, would be to compute p as J(S) and project it onto Y to obtain $\pi_Y(p)$:

$$\pi_Y(p) = \pi_Y(J(S)).$$

This may be much more expensive than is necessary. To give an extreme example, let $X = \{\{v_1, v_2, v_3\}, \{v_2, v_3, v_4\}, \{v_1, v_3, v_4\}\}$ and $Y = \{\{v_1, v_2\}, \{v_3, v_4\}\}$. From Proposition 2, $\pi_Y(p)$ is relatively quickly obtained by projection from two of the elements of S:

$$\pi_Y(p) = \pi_{\{\{v_1, v_2\}, \{v_3, v_4\}\}}(p))$$
$$= \{\pi_{\{v_1, v_2\}}(\pi_{\{v_1, v_2, v_3\}}(p)), \pi_{\{v_3, v_4\}}(\pi_{\{v_2, v_3, v_4\}}(p))\}.$$

Computing $\pi_Y(p)$ as $\pi_Y(J(S))$ requires an expensive iterative join procedure (Cavallo and Klir, 1981). This illustrates the general principle that whenever

$Y \subseteqq X$, $\pi_Y(p)$ is most efficiently computed as

$$\pi_Y(p) = \{\pi_{V_i}(\pi_{A_i}(p)), \ldots, \pi_{V_j}(\pi_{A_j}(p))\},$$

where $Y = \{V_i, \ldots, V_j\}$, $V_i \subseteqq A_i, \ldots, V_j \subseteqq A_j$, and $A_i, \ldots, A_j \in X$.

Consider now a situation in which $p = J(\pi_X(p))$, $\pi_X(p)$ is given, and $\pi_{V_0}(p)$ is desired, where $V_0 \subseteqq V$, the scheme for p. If $V_0 \subseteqq A \in X$, then $\pi_{V_0}(p) = \pi_{V_0}(\pi_A(p))$, from Proposition 2. Otherwise, there are two approaches possible.

One could compute $\pi_{V_0}(p)$ as $\pi_{V_0}(J(\pi_X(p)))$, as explained above, but this may be unnecessarily expensive.

For loopless (α-acyclic) structures (Pittarelli, 1990; Section 3) it may be possible to join over a much smaller structure and then, if necessary, project the resulting distribution onto the set V_0. Any structure $X = \{V_1, \ldots, V_m\}$ (loopless or not) may be partitioned into a set of connected components $X_C = \{E_1, \ldots, E_k\}$ such that:

(1) $\bigcup_{V \in E_i} V \cap \bigcup_{V \in E_j} V = \varnothing$ for all $i, j \in \{1, \ldots, k\}$.

(2) For any $i \in \{1, \ldots, k\}$ V_l, $V_j \in E_i$ implies that there is a sequence $V_{i_1}, \ldots, V_{i_n}$ such that

(*a*) $V_{i_q} \in E_i$, $q \in \{1, \ldots, n\}$,

(*b*) $V_{i_q} \cap V_{i_{q+1}} \neq \varnothing$, $q \in \{1, \ldots, n-1\}$,

(*c*) $V_{i_1} = V_l$, $V_{i_n} = V_j$.

Let X_{V_0} denote the subset of members of X contained in some connected component involving elements of V_0:

$$X_{V_0} = \{\bigcup_{C \in X_C} | \text{for some } V_i \in C,\ V_i \cap V_0 \neq \varnothing\}.$$

For a loopless structure $X = \{V_1, \ldots, V_m\}$, let $\Sigma(X, V_0) = (V_{i_1}, \ldots, V_{i_n})$ denote the reverse of the sequence in which elements $V_i \in X_{V_0}$ are eliminated by the algorithm (*see* Algorithm 3.1 in Pittarelli, 1990; this issue):

$k := m$;
$\Sigma(X, V_0) := \varnothing$;
$W := X_{V_0}$;
repeat in any order until $W = \{\varnothing\}$

(1) if v appears in only one $V_{i_w} \in W$ then $V_{i_W} := V_{i_W} - \{v\}$

(2) if $V_{i_W} \subseteqq V_{j_W}$, where $i \neq j$,
then begin
$W := W - \{V_{i_W}\}$;

$\Sigma(X, V_0)_k := V_{i_w}$;
$k := k-1$
end.

Let $\sigma(X, V_0)$ denote the shortest subsequence of $\Sigma(X, V_0)$ such that no element of $\Sigma(X, V_0) - \sigma(X, V_0)$ includes an element of V_0. A (pairwise) join expression (*see* Pittarelli, 1990; this issue) is constructed from $\sigma(X, V_0) = (V_{\sigma_1}, \ldots, V_{\sigma_j})$ as

$$p_1(V_{\sigma_1}) \times p_2(V_{\sigma_2} - V_{\sigma_1} | V_{\sigma_1} \cap V_{\sigma_2}) \times \ldots \times p_n(V_{\sigma_n} - V_{\sigma_{n-1}} - \ldots - V_{\sigma_1} | V_{\sigma_n} \cap (V_{\sigma_{n-1}} \cup \ldots \cup V_{\sigma_1})$$

where, e. g., $p_1(V_{\sigma_1})$ abbreviates $\pi_{V_{\sigma_1}}(p)$. The resulting probability distribution is

$$\pi_{V_{\sigma_1} \cup \ldots \cup V_{\sigma_j}}(p) = J(\pi_{\sigma(X, V_0)}(p)).$$

By Proposition 1, projection of this distribution onto the set V_0 yields $\pi_{V_0}(p)$, which coincides with $\pi_{V_0}(J(\pi_X(p)))$. Therefore,

Theorem 2. — $\pi_{V_0}(J(\pi_X(p))) = \pi_{V_0}(J(\pi_{\sigma(X, V_0)}(p)))$. □

Since $\sigma(X, V_0) \subseteqq X$ and the refinement algorithm and projection are less expensive than the join operation, this provides a more efficient method of computing $\pi_{V_0}(p)$ than as $\pi_{V_0}(J(\pi_X(p)))$.

Example 3.1. — $p = J(\pi_X(p))$, where $X = \{\{v_1, v_2\}, \{v_3, v_4\}, \{v_5, v_6\}, \{v_2, v_7\}\}$. Let $V_0 = \{v_1, v_7\}$. The structure system $\pi_X(p)$ is

v_1	v_2	$\pi_{\{v_1, v_2\}}(p)(.)$	v_3	v_4	$\pi_{\{v_3, v_4\}}(p)(.)$
0	0	0.2	0	0	0
0	1	0.3	0	1	0.5
1	0	0.4	1	0	0.25
1	1	0.1	1	1	0.25

v_5	v_6	$\pi_{\{v_5, v_6\}}(p)(.)$	v_2	v_7	$\pi_{\{v_2, v_7\}}(p)(.)$
0	0	0.3	0	0	0.1
0	1	0.3	0	1	0.4
1	0	0.3	1	0	0.4
1	1	0.1	1	1	0.1

For this example,

$$X_C = \{\{\{v_1, v_2\}, \{v_2, v_7\}\}, \{\{v_3, v_4\}\}, \{\{v_5, v_6\}\}\}$$

and

$$X_{V_0} = \{\{v_1, v_2\}, \{v_2, v_7\}\}.$$

Independent of elimination order, $\sigma(X, V_0) = X_{V_0}$ in this case. Thus,

$$\pi_{\{v_1, v_7\}}(p) = \pi_{\{v_1, v_7\}}(J(\pi_{\{v_1, v_2\}}(p), \pi_{\{v_2, v_7\}}(p))) :$$

v_1	v_7	$\pi_{\{v_1, v_7\}}(p)(.)$
0	0	0.292
0	1	0.208
1	0	0.208
1	1	0.292

Notice that this distribution does not equal the join of the projections onto the individual variables v_1 and v_7:

v_1	v_7	$J(\{\pi_{\{v_1\}}(p), \pi_{\{v_7\}}(p)\})(.)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

In particular, notice that its entropy is lower. Information, quantified as negative entropy, is lost by first projecting and then joining. This relation holds in general, regardless of the reconstructability of p from its projections onto X. Let Y be any refinement of X that is a cover of V_0.

THEOREM 3. – $H(\pi_{V_0}(J(\pi_X(p)))) \leqq H(J(\pi_Y(\pi_X(p))))$.

Proof. – Since $Y \leqq X$, it follows from Proposition 10 that

$$E_V(\pi_X(p)) \leqq E_V(\pi_Y(p)).$$

Therefore,

$$\pi_{V_0}(E_V(\pi_X(p))) \leqq \pi_{V_0}(E_V(\pi_Y(p))).$$

By Proposition 7,

$$\pi_{V_0}(E_V(\pi_Y(p))) = E_{V_0}(\pi_Y(p)) = E(\pi_Y(p)).$$

$$J(\pi_X(p)) \in E_V(\pi_X(p)).$$

Therefore,

$$\pi_{V_0}(J(\pi_X(p))) \in \pi_{V_0}(E_V(\pi_X((p)))).$$

Since $\pi_{V_0}(E_V(\pi_X(p))) \leqq E(\pi_Y(p))$, it follows that

$$\pi_{V_0}(J(\pi_X(p))) \in E(\pi_Y(p)).$$

By Theorem 1, $E(\pi_Y(p)) = E(\pi_Y(\pi_X(p)))$. By definition, $J(\pi_Y(\pi_X(p)))$ is the maximum entropy element of $E(\pi_Y(\pi_X(p)))$. So,

$$H(\pi_{V_0}(J(\pi_X(p)))) \leqq H(J(\pi_Y(\pi_X(p))))). \quad \square$$

Example 3.1 illustrates the special case of Theorem 3 in which $p = J(\pi_X(p))$ and $Y = \{\{v\} \mid v \in V_0\}$. A more important special case is $V_0 = V$ and $X = \{V\}$, for an arbitrary distribution p :

$$H(\pi_V(J(\pi_{\{V\}}(p)))) \leqq H(J(\pi_Y(\pi_{\{V\}}(p))))$$

becomes, by Proposition 1 and Theorem 1,

$$H(p) \leqq H(J(\pi_Y(p))).$$

This is a statement of the standard reconstruction problem of reconstructability analysis, in which a known p is projected onto a structure Y and the difference $H(J(\pi_Y(p))) - H(p)$, which in general is nonzero, measures the degree to which p is reconstructable from its projections onto Y (Cavallo and Klir, 1981; Higashi, 1984).

Although all members of a reconstruction family $E(\pi_X(p))$ are by definition equivalent with respect to projections onto $X = \{V_1, \ldots, V_m\}$ and thus onto elements of X, they are not necessarily equivalent with respect to projections onto arbitrary sets $V_0 \leqq V_1 \cup \ldots \cup V_m$. When this equivalence does hold, however, what is manifested is a generalization of the standard notion of the identifiability of a distribution p from its projection onto a structure X to the identifiability of $\pi_{V_0}(p)$ from the projection of p onto X:

$$\pi_{V_0}(E_V(\pi_X(p))) = \{\pi_{V_0}(p)\}.$$

When $V_0 = V$, the scheme for p, this reduces, by Proposition 8, to the ordinary identifiability concept:

$$E(\pi_X(p)) = \{p\}.$$

[A similar generalization of reconstructability,

$$\pi_{V_0}(p) = J(\pi_Y(\pi_{V_0}(p))),$$

where Y is a model of V_0, has as an analogue in relational database theory the notion of an *embedded join dependency* (Fagin and Vardi, 1986).] Because

projections $\pi_{V_0}(p)$ are not in general identifiable from the projection of p onto an arbitrary structure, if one has as data a structure system S, which may be viewed as $\pi_X(p)$ for some unknown p, and information is required in the form of a distribution over some set V_0, it will not do to form $\pi_{V_0}(J(S))$. For the unknown actual distribution p (which needn't be reconstructable from X), it may be that $\pi_{V_0}(p)$ is very different from $\pi_{V_0}(J(S))$. To illustrate, for the structure system

G	H	$p_1(.)$	H	I	$p_2(.)$
0	0	0.20	0	0	0.15
0	1	0.35	0	1	0.35
1	0	0.30	1	0	0.25
1	1	0.15	1	1	0.25

$\min_{p \in E(\{p_1, p_2\})} p(GI=11)=0.15$, $\max_{p \in E(\{p_1, p_2\})} p(GI=11)=0.45$. When the projection onto the set V_0 of variables of actual interest is not identifiable from a structure system (which can be determined via linear programming), it seems advisable to work with the set of distributions $\pi_{V_0}(E(S))$, which is guaranteed to contain $\pi_{V_0}(p)$. Methods for utilizing such sets are discussed next.

A structure system may represent data from studies undertaken on multiple sets of variables, independently or in some coordinated manner. To illustrate the latter, consider a surveillance problem in which one wishes to determine the relative frequencies with which a moving object occupies the cells of a 3-dimensional grid over some period of time, but one is able to observe motion in only two of the planes (since, e. g., sensors mounted in such a way as to detect movement in the third plane directly might be damaged, or might be detected by the object, causing it to change its normal pattern of movement, etc.). Bisecting each dimension (X, Y, Z) into lower ("L") and upper ("U") halves, observed relative frequencies of occupation of 2-dimensional cells might be represented as the structure system $S=\{\pi_{\{X,Y\}}(p), \pi_{\{X,Z\}}(p)\}$ below, where p over scheme $\{X, Y, Z\}$ is the actual but unknown relative frequency distribution over the 3-dimensional space:

X	Y	$\pi_{\{X,Y\}}(p)(.)$	X	Z	$\pi_{\{X,Z\}}(p)(.)$
L	L	0.1	L	L	0.15
L	U	0.1	L	U	0.05
U	L	0.4	U	L	0.3
U	U	0.4	U	U	0.5

There is no reason to believe in this case that $p=J(S)$ or, for that matter, any other one in particular of the infinitely many elements of E(S). Traditionally, for purposes, say, of decision analysis, it has been the practice to estimate p as J(S) and then apply standard Bayesian techniques. As argued elsewhere (Pittarelli, 1989), not only is this unnecessary, but it can produce misleading results.

Example 3.2. – To illustrate, consider a decision problem utilizing the information in S, above. It is necessary to guess which cell the object is currently occupying, with decision matrix

	$s_{XYZ=LLL}$	s_{LLU}	s_{LUL}	s_{LUU}	s_{ULL}	s_{ULU}	s_{UUL}	s_{UUU}
a_{LLL}	100	0	0	0	0	0	0	−50
a_{LLU}	0	010	0	0	0	0	−10	0
a_{LUL}	0	0	10	0	0	−5	0	0
a_{LUU}	0	0	0	10	−5	0	0	0
a_{ULL}	0	0	0	−5	10	0	0	0
a_{ULU}	0	0	−5	0	0	10	0	0
a_{UUL}	0	−10	0	0	0	0	10	0
a_{UUU}	−30	0	0	0	0	0	0	100

where a_t and s_t denote, respectively, the act of guessing that, and the state in which, the object is occupying cell t, and the entry in row a_t, column $s_{t'}$ is the utility $u(a_t, s_{t'})$ of act a_t when the object is in cell t'.

If one knew that $p=p^*=J(S)$, one would compute p^* and select an act a^* maximizing expected utility:

$$\sum_t u(a^*, s_t)\, p^*(s_t) = \max_a \sum_t u(a, s_t)\, p^*(s_t).$$

For the example, the act a_{UUU} uniquely maximizes expected utility under the estimate p^* : $\sum_t u(a_{UUU}, s_t)\, p^*(s_t) = 22.75$. For the data S, however, it is possible to determine that this act maximizes expected utility relative to *any* $p \in E(S)$. Using linear programming, *expected utility intervals* U(a) can be calculated (Seidenfeld, 1983; Pittarelli, 1988, 1989) for each act. Relative to a given distribution p, let $e(a)=\sum_t u(a, s_t)\, p(s_t)$. Then, relative to a structure system S,

$$U(a) = [\min_{p \in E(S)} e(a), \max_{p \in E(S)} e(a)].$$

For this decision problem, the utility intervals are:

$$U(a_{LLL}) = [-15, 5]$$
$$U(a_{LLU}) = [-3, 0.5]$$
$$U(a_{LUL}) = [-1.5, 0.5]$$
$$U(a_{LUU}) = [-1.5, 0.5]$$
$$U(a_{ULL}) = [-0.25, 3]$$
$$U(a_{ULU}) = [-0.5, 3.75]$$
$$U(a_{UUL}) = [-0.5, 0.3]$$
$$U(a_{UUU}) = [7, 38.5].$$

Act a_{UUU} *dominates* each of the others in the sense that $\min U(a_{UUU}) > \max U(a)$, for all other acts a. There will not always be a unique dominating action. There exist techniques applicable in such cases, but they are somewhat controversial; e. g., maximizing the minimum expected utility (Pittarelli, 1988, 1989). The smaller the set E(S) in a decision problem with probability data in the form of a structure system, the narrower the utility intervals U(*a*) and hence the greater the likelihood that a unique dominating act will be identified. This motivates the discussion of the data synthesis optimization procedures to be presented next, which take as input structure systems and produce as output sets of probability distributions.

Given a structure system S with structure $X = \{V_1, \ldots, V_m\}$, suppose that only variables $V_0 \subseteqq V = V_1 \cup \ldots \cup V_m$ are of interest for some purpose (e. g., a decision problem). With p the actual but unknown distribution such that $S = \pi_X(p)$, the probability distribution $p_0 = \pi_{V_0}(p)$ is desired. Unlike the type of problem discussed previously, in which p is recontructable from its projections onto X, it will not necessarily be the case that $\pi_{V_0}(p) = \pi_{V_0}(J(S))$. Thus it will not in general be possible to construct p_0 itself from S. Instead, a set (convex polyhedron) of distributions containing p_0 is determined. This set can be used in decision making (as illustrated above), or bounds on components $p_0(t)$ can be calculated from it using linear programming, etc.

There are trivial cases in which $p_0 = \pi_{V_0}(p)$ is determinable. If $V_0 \in X$, then p_0 is immediately given as an element of S. If $V_0 \subset V_k \in X$, then, from Proposition 2, $p_0 = \pi_{V_0}(\pi_{V_k}(p))$.

Otherwise, constraints determining a set containing p_0 must be pieced together from different elements of S. There are two basic approaches. In the first, S is extended and the resulting set of distributions is projected onto

V_0. Since, by Proposition 3, $p \in E(\pi_X(p))$, it follows that

$$p_0 \in \pi_{V_0}(E(\pi_X(p))).$$

Alternatively, projection may occur before extension. The extreme case would entail forming individual distributions $\pi_{\{v\}}(p)$ for each $v \in V_0$ and taking their extension. Let $X_0 = \{\{v\} \mid v \in V_0\}$. Then, by reasoning similar to that in the proof of Theorem 3, $p_0 \in E_{V_0}(\pi_{X_0}(p)) = E(\pi_{X_0}(S))$. Less extreme would be to form the structure system $\pi_{X'}(p)$, where $X' = \{V_0 \cap V_i \mid V_i \in X, V_0 \cap V_i \neq \varnothing\}$, and extend it. It is straightforward to show that $p_0 \in E(\pi_{X'}(p))$ also. At slightly greater expense (more linear equations, but the same number of variables, to characterize the set of distributions), this method gives more information regarding p_0: from Proposition 10, since $X_0 \leqq X'$, $E(\pi_{X'}(p)) \subseteqq E(\pi_{X_0}(p))$.

Although they are more efficient, neither of the "project first" techniques yields constraints on p_0 as strong as those determinable by extending S and then projecting onto V_0.

Theorem 4. – $\pi_{V_0}(E(\pi_X(p))) \subseteqq E(\pi_{X'}(\pi_X(p))) = E(\pi_{X'}(p))$.

Proof. – Since $X' \leqq X$, it follows from Proposition 10 that

$$E_V(\pi_X(p)) \subseteqq E_V(\pi_{X'}(p)),$$

which implies that

$$\pi_{V_0}(E_V(\pi_X(p))) \subseteqq \pi_{V_0}(E_V(\pi_{X'}(p))),$$

which, by Proposition 7, implies

$$\pi_{V_0}(E_V(\pi_X(p))) \subseteqq E_{V_0}(\pi_{X'}(p)),$$

i. e., that

$$\pi_{V_0}(E(\pi_X(p))) \subseteqq E(\pi_{X'}(p)). \quad \square$$

[Therefore, $\pi_{V_0}(E(\pi_X(p))) \subseteqq E(\pi_{X_0}(p))$ also.]

X_0 and X are, respectively, the most and least refined structures over which projections of p are available that cover some $V \supseteqq V_0$. For any such covering structure W, $X_0 \leqq W$. If also $W \leqq X$, then Proposition 10 guarantees that $p_0 \in \pi_{V_0}(E(\pi_W(p)))$. The smallest of these sets guaranteed to contain p_0 is $\pi_{V_0}(E(\pi_X(p)))$. Note, however, that less refined structures W do not necess-

arily determine strictly smaller sets $\pi_{V_0}(E(\pi_W(p)))$. For p defined as

E	F	G	$p(.)$
0	0	0	0.0
0	0	1	0.0
0	1	0	0.2
0	1	1	0.2
1	0	0	0.0
1	0	1	0.0
1	1	0	0.4
1	1	1	0.2

and with $V_0 = \{E, G\}$, $p_0 = \pi_{\{E, G\}}(p)$ is identifiable from the most refined covering model, $X_0 = \{\{E\}, \{G\}\}$. Therefore, nothing is to be gained by considering less refined models:

$$\pi_{V_0}(E(\pi_W(p))) = \{\pi_{V_0}(p)\}, \quad \text{for any } X_0 \leqq W.$$

Example 3.3. – As an application of the "extend, then project" principle of Theorem 4, consider a decision problem utilizing the data of Example 3.1, with decision matrix

	$s_{v_1 v_7 = 00}$	s_{01}	s_{10}	s_{11}
a_1	50	0	−5	1,000
a_2	0	10	20	0
a_3	400	0	0	10

The set of variables of actual interest is $V_0 = \{v_1, v_7\}$. Data are available in the form of a structure system $\pi_X(p)$, where $X = \{\{v_1, v_2\}, \{v_3, v_4\}, \{v_5, v_6\}, \{v_2, v_7\}\}$, and p is unknown and unidentifiable. The least expensive solution based on extension and not estimation of $\pi_{V_0}(p)$ utilizes projections onto $X_0 = \{\{v_1\}, \{v_7\}\}$:

v_1	$\pi_{\{v_1\}}(p)(.)$	v_7	$\pi_{\{s_7\}}(p)(.)$
0	0.5	0	0.5
1	0.5	1	0.5

Utility intervals calculated from this structure system are indecisive:

$$U(a_1) = [-2.5, 525]$$
$$U(a_2) = [0, 15]$$
$$U(a_3) = [0, 205].$$

Extending the entire structure system and then projecting onto V_0 yields the utility intervals

$$U(a_1) = [208, 525]$$
$$U(a_2) = [0, 9]$$
$$U(a_3) = [82, 205],$$

from which it can be determined that a_1 is uniquely optimal with respect to expected utility maximization: for any $p' \in \pi_X(p)$, act a_1 uniquely maximizes expected utility relative to $\pi_{V_0}(p')$.

The wider (and indecisive) "project first" utility intervals are less expensive to calculate, requiring a linear program involving only 4 unknowns per endpoint, vs. 128 unknowns for the "extend first" method. The theorem below provides a method ("refine, then extend, then project") for determining the narrow "extend first" intervals much more efficiently (8 unknowns).

Theorem 5 (Pittarelli, 1988). – $\pi_{V_0}(E(\pi_X(p))) = \pi_{V_0}(E(\pi_W(p)))$, where W is the structure resulting from application of the algorithm:

(1) $W \leftarrow X$.

(2) Repeat in any order until neither has any effect on the current value of W:

(*a*) if a variable $v \notin V_0$ appears in only one element of W, remove v from that element,

(*b*) if W contains elements V_i and V_j such that $V_i \subset V_j$, then $W \leftarrow W - \{V_i\}$. □

(Observe that W does not in general coincide with $\sigma(X, V_0)$.)

Since $W \leqq X$, characterization of $\pi_{V_0}(E(\pi_W(p)))$ requires fewer linear equations in fewer unknowns. For the X and V_0 of Example 3.3, $W = \{\{v_1, v_2\}, \{v_2, v_7\}\}$. Thus, the utility intervals calculated by extending $\pi_X(p)$ and then projecting onto V_0 can be calculated by first refining X to W (by means of the polynomial-time algorithm of Theorem 5) before extending and then projecting onto V_0; *i.e.*, they can be determined from the structure system

v_1	v_2	$\pi_{\{v_1, v_2\}}(p)(.)$	v_2	v_7	$\pi_{\{v_1, v_7\}}(p)(.)$
0	0	0.2	0	0	0.2
0	1	0.3	0	1	0.4
1	0	0.4	1	0	0.3
1	1	0.1	1	1	0.1

4. Conclusion and suggestions for future research

Structure systems are treated as probabilistic databases from which information in various forms may be extracted. The algebra of standard recon-

structability analysis is extended slightly, and a few theorems relevant to procedures for data synthesis are presented.

Although these procedures are illustrated through application to problems of decision analysis, their scope is much wider. It would therefore be worthwhile not only for its mathematical interest, but also for pragmatic reasons, to extend further the rudimentary algebra of probabilistic structure systems sketched in this paper.

Acknowledgements

This line of research originated in discussions with Roger Cavallo, to whom I am intellectually deeply indebted. I would also like to thank Philip Meguire for his translation into French of the abstract and for many years of discussion and correspondence that have influenced me greatly.

References

N. BOURBAKI, *Théorie des Ensembles*. Hermann C^ie^, Paris, 1954.

R. CAVALLO and G. KLIR, Reconstructability analysis of multi-dimensional relations: a theoretical basis for computer-aided determination of acceptable systems models, *Int. J. of General Systems*, 5, 1979, pp. 143-171.

R. CAVALLO and G. KLIR, Reconstructability analysis: evaluation of reconstruction hypotheses, *Int. J. of General Systems*, 7, 1981, pp. 7-32.

R. CAVALLO and M. PITTARELLI, The theory of probabilistic databases. *Proc. 13th Int. Conf. on Very Large Databases (VLDB)*, 1987, pp. 71-81.

R. FAGIN and M. VARDI, The theory of data dependencies – a survey, *in:* M. ANSHEL and W. GEWIRTZ, Eds., *Mathematics of Information Processing*, American Mathematical Society, Providence, Rhode Island, 1986, pp. 19-72.

M. HIGASHI, A systems modelling methodology: probabilistic and possibilistic approaches, *Ph. D. dissertation*, SUNY-Binghamton, Binghamton, New York, 1984.

E. T. JAYNES, Prior information and ambiguity in inverse problems, *in:* D. McLAUGHLIN, Ed., *Inverse problems*, SIAM-AMS Proceedings, 14, American Mathematical Society, Providence, Rhode Island, 1984, pp. 151-166.

P. M. LEWIS, Approximating probability distributions to reduce storage requirements, *Information and Control*, 2, 1959, pp. 214-225.

M. PITTARELLI, Decision making with linear constraints on probabilities, *Proc. 4th Workshop on Uncertainty in Artifical Intelligence*, 1988, pp. 283-290.

M. PITTARELLI, Uncertainty and estimation in reconstructability analysis, *Int. J. of General Systems*, 15, 1989, pp. 1-58.

M. PITTARELLI, Reconstructability analysis: an overview, *Revue Int. de Systemique*, this issue, 1990.

T. SEIDENFELD, Decisions with indeterminate probabilities, *The Behavioral and Brain Sciences*, 2, 1983, pp. 259-261.

REVUE INTERNATIONALE DE SYSTEMIQUE
Vol. 4, N° 1, 1990, pp. 65 à 78

RECONSTRUCTION PRINCIPLE OF INDUCTIVE REASONING

George J. KLIR

State University of New York [1]

Abstract

A new principle of inductive reasoning, which is based upon reconstructability analysis, is discussed. The principle differs from the straight rule, which is usually associated with inductive reasoning. Experimental studies are described by which the principle is confirmed and its domain of applicability partially delimited. The connection of the principle to the notion of pragmatic information is also mentioned.

It is assumed that the reader is familiar with the reconstruction problem of reconstructability analysis, which is overviewed in this issue by Pittarelli (1990).

Résumé

Nous présentons un nouveau principe de raisonnement inductif fondé sur l'analyse de la reconstructibilité. Ce principe se distingue de la règle simple, assimilée d'habitude au raisonnement inductif. Nous traitons d'études expérimentales qui confirment notre principe et qui délimitent en partie son domaine d'application. Nous signalons le rapport entre notre principe et la notion d'information pragmatique.

Nous supposons que le lecteur connaît le problème de reconstruction tel qu'il se présente dans l'analyse de la reconstructibilité, résumée dans ce numéro par Pittarelli (1990).

1. Introduction

During my study of the reconstruction problem (Klir, 1976, 1984, 1986), one of the two problems addressed by reconstructability analysis, I made a

1. Binghamton, New York 13901, U.S.A. Department of Systems Science, Thomas J. Watson School of Engineering, Applied Science and Technology.

Rev. intern. systémique. 0980-1472 Vol. 4/90/01/ 65 /14/$ 3.40/