PROBABILISTIC MODELS IN ARTIFICIAL INTELLIGENCE

Jiří Vomlel

18 September 1995

Department of Decision-Making Theory
Institute of Information Theory and Automation
Pod vodárenskou věží 4,
182 08 Praha 8, Czech Republic

Department of Control Engineering
Faculty of Electrical Engineering
Czech Technical University
Technická 2,
160 00 Praha 6, Czech Republic

18 September 1995

Department of Decision-Making Theory
Institute of Information Theory and Automation
Pod vodárenskou věží 4,
182 08 Praha 8, Czech Republic

Department of Control Engineering
Faculty of Electrical Engineering
Czech Technical University
Technická 2,
160 00 Praha 6, Czech Republic

1 Introduction

This article is devoted to Probabilistic Models used in the field of Artificial Intelligence. A Probabilistic Model is one from possibilities for handling uncertain information in AI. In its frame integration of knowledge can be defined as building a joint probability distribution from a set of oligodimensional distributions. We can use the well known *Iterative Proportional Fitting* (IPF) procedure. It is proved that IPF converges if the initial distribution is uniform distribution and if the set of oligodimensional distributions is consistent, i.e. if there exists at least one distribution with the given marginals [Csis 75]. If the result distribution is decomposable then it can be effectively stored and handled so that it make possible fast computation of conditional probabilities [Háj,Havr,Jir 92].

In addition Csiszár showed that IPF converges to the distribution that maximizes entropy from all distributions having input oligodimensional distributions as their marginals. Some reasons for the maximal entropy principle are given in [Jir 92, Per 83].

In the case of an inconsistent input set of oligodistributions the IPF procedure tends to cycle. Nevertheless, we wish to get some representative of the input knowledge even in case when it is inconsistent. In section 4 there is short discussion how it can happen that from our consistent world we derive an inconsistent input set. Therefore, we endeavor to find a modification of IPF which converges even in inconsistent situations, the marginals of the resulting distribution are somehow close to the input distributions and independent of the ordering of the input oligodimensional distributions. Two modifications of IPF were described in [Jir,Vom 95]. In this article both these modifications are also described together with several examples. First one is mentioned in Section 6 where its convergence is proven. The second one is described in Section 7.

There are plenty of possibilities how to create input oligodimensional distribution sets from given data vectors. This is typically an operation which needs expert knowledge. An approach searching for the best approximation (in a sense) of data is mentioned in Section 9.

2 Probabilistic models

Let V denote a finite index set of a system $\{X_i\}_{i\in V}$ of finite-valued variables for which we are looking for a joint probability distribution $Q((X_i)_{i\in V})$. For each $i \in V$, $\mathcal{X}_i$ will denote a finite set of values of the variable X_i. Further, let C_k for $k = 1, \ldots, K$ be subsets of this index set V. For each C_k we assume that a probability distribution P_k defined for variables $(X_i)_{i\in C_k}$ is given. This system of oligodimensional distributions will be denoted by Θ:

$$\Theta = \{P_1, \ldots, P_K\}.$$

Considering any joint probability distribution $Q((X_i)_{i\in V})$ and a subset $C \subseteq V$, the symbol Q^C denotes the marginal distribution of Q defined for variables $(X_i)_{i\in C}$. For any system of oligodimensional distributions $\Theta = \{P_1, \ldots, P_K\}$ defined for variables $\{X_i\}_{i\in C_1}, \ldots, \{X_i\}_{i\in C_K}$ respectively we can specify the system

$$\Pi(\Theta)=\{R((X_i)_{i\in V}) : R^{C_k} = P_k, \forall k = 1, \ldots, K\}$$

of all the joint probability distributions consistent with all the oligodimensional distributions from Θ.

In the next sections we will use following definitions.

Definition 1 (Entropy) $H(P)$ denotes the *Shanon entropy* of the distribution $P((X_i)_{i\in V})$ defined

$$H(P) = -\sum_{x\in \times_{i\in V}\mathcal{X}_i} P(x)\cdot \log P(x).$$

Definition 2 (Variance) $V(P,R)$ denotes the *Total Variance* of distributions $P((X_i)_{i\in V})$ and $R((X_i)_{i\in V})$ defined

$$V(P,R) = \sum_{x\in \times_{i\in V}\mathcal{X}_i} \mid P(x) - R(x) \mid$$

Definition 3 (I-divergence) $I(P \parallel R)$ denotes the *Kullback-Leibler divergence (I-divergence)* of distributions $P((X_i)_{i\in V})$ and $R((X_i)_{i\in V})$ defined

$$I(P \parallel R) = \sum_{x\in \times_{i\in V}\mathcal{X}_i} P(x)\cdot \log\frac{P(x)}{R(x)}$$

in case that the expresion is finite. Otherwise

$$I(P \parallel R) = +\infty.$$

Definition 4 (I-projection) Distribution Q is an I-projection of a distribution S into the class of joint distribution

$$\Pi(\{P_j((X_i)_{i\in C_j})\}) = \{R((X_i)_{i\in V}) : R^{C_j} = P_j\})$$

if

$$Q \in \Pi(\{P_j\}) \quad \& \quad I(Q \parallel S) = \min_{R\in\Pi(\{P_j\})} I(R \parallel S).$$

Definition 5 (Informational content) $I(P)$ denotes the *Informational content* of the distribution $P((X_i)_{i\in V})$ defined

$$I(P) = \sum_{x\in \times_{i\in V}\mathcal{X}_i} P(x)\cdot \log\frac{P(x)}{\prod\limits_{k\in V} P^{\{k\}}(x)},$$

where $P^{\{k\}}(x)$ denotes the one-dimensional marginal distribution for the variable X_k i.e. $P(X_k)$.

Definition 6 (Strong consistency) The system $\Theta = \{P_1, \ldots, P_k\}$ is *strongly consistent* (further we shall use also *consistent*) if $(\Pi(\Theta) \neq \emptyset)$.

Definition 7 (Weak consistency) We shall say that a system $\Theta = \{P_1, \ldots, P_k\}$ is *weakly consistent* if for any couple of different indices $k, \ell \in \{1, \ldots, K\}$

$$P_k((X_i)_{i\in(C_k\cap C_\ell)}) = P_\ell((X_i)_{i\in(C_k\cap C_\ell)})$$

or, equivalently

$$P_k^{C_k\cap C_\ell} = P_\ell^{C_k\cap C_\ell}.$$

3 Tests of consistency

So far no general procedure for consistency tests is known. In this section we will describe a fast test for a particular case. This approach is similar to [Kell 64].

Let us start with definition of Running Intersection Property.

Definition 8 (Running Intersection Property) Let C_{j_k} for $k = 1, \ldots, K$ be an ordering of sets $C_1, \ldots, C_K$. This ordering meets the *Running Intersection Property (RIP)* if

$$\forall l = 2, \ldots, K \quad \exists k \quad (1 \leq k < l) \quad ((C_l \cap \bigcup_{m=1}^{l-1} C_m) \subset C_k).$$

In the case of input set of distributions satisfying RIP we can very simplify test of *weak* and consequently also *strong consistency* of input set. Following theorem show how it could be done.

Theorem 1 (Simplification of weak consistency test) *Let*

$$\Theta = \{P_1((X_i)_{i\in C_1}), \ldots, P_K((X_i)_{i\in C_K})\}$$

be a system of distributions and C_{j_k} for $k = 1, \ldots, K$ be an ordering of sets $C_1, \ldots, C_K$ meeting the RIP*, then following two statements are equivalent:*

(a) *system Θ is weakly consistent*

(b) *system Θ is consistent*

(c) $\forall k = 2, \ldots, K \quad \exists \ell \quad (1 \leq \ell < k \quad \& \quad (C_k \cap \bigcup_{m=1}^{k-1} C_m) \subset C_\ell \quad \& \quad P_k^{C_k\cap C_\ell} = P_\ell^{C_k\cap C_\ell})$.

Proof. **(a) $\Rightarrow$ (c)** If system is weakly consistent then, according to definition, for any couple of different indices

$$k, \ell \in \{1, \ldots, K\} \quad (P_k^{C_k\cap C_\ell} = P_\ell^{C_k\cap C_\ell}).$$

therefore this equality must be valid also for a subset of all couples defined by (c).

(c) $\Rightarrow$ (a) Condition of weak consistency of system Θ can be rewritten :

$$\forall k = 2, \ldots, K \quad \forall j < k \quad (P_k^{C_k\cap C_j} = P_j^{C_k\cap C_j}). \tag{1}$$

If $k = 2$ than it is trivial to show that if $j = 1$ satisfy statement (c) then also statement (a) is satisfied. It is because the condition

$$P_2^{C_2\cap C_1} = P_1^{C_2\cap C_1}$$

is the same in both statements.
Suppose that the theorem is valid for $k \leq i$. We want to show it must be valid also for $k \leq i+1$. In other words we want to prove that

$$\forall k = 2, \ldots, i+1 \quad \forall j < k \quad (P_k^{C_k \cap C_j} = P_j^{C_k \cap C_j})$$

is true. If the theorem is valid for $k \leq i$ than following holds: From (1) we can write:

$$\forall k = 2, \ldots, i \quad \forall j < k \quad (P_k^{C_k \cap C_j} = P_j^{C_k \cap C_j}). \tag{2}$$

From statement (c) we can write:

$$\text{for} \quad k = i+1 \quad \exists \ell \quad (1 \leq \ell < k) \quad \&$$

$$(C_k \cap \bigcup_{m=1}^{k-1} C_m) \subset C_\ell \quad \& \tag{3}$$

$$P_k^{C_k \cap C_\ell} = P_l^{C_k \cap C_\ell}. \tag{4}$$

At first let us suppose that $j \neq \ell$. Because $j < k = i+1$ either $j < i$ or $l < i$ and therefore for $k = i+1$ from (2) follows:

$$P_j^{C_j \cap C_\ell} = P_\ell^{C_j \cap C_\ell}. \tag{5}$$

According to (3) $(C_{i+1} \cap C_j) \subset C_\ell$ and therefore we can marginalize both sides of (5) than we get

$$P_j^{C_j \cap C_{i+1}} = P_{i+1}^{C_j \cap C_{i+1}}. \tag{6}$$

Equation (4) we can rewrite for given $k = i+1$

$$P_{i+1}^{C_{i+1} \cap C_\ell} = P_\ell^{C_{i+1} \cap C_\ell} \tag{7}$$

and again marginalize

$$P_{i+1}^{C_{i+1} \cap C_j} = P_\ell^{C_{i+1} \cap C_j}. \tag{8}$$

Comparing (6) and (8) we receive

$$P_{i+1}^{C_{i+1} \cap C_j} = P_j^{C_{i+1} \cap C_j}. \tag{9}$$

Secondly we suppose that $j = \ell$. In this case we receive equation (9) directly by putting $j = \ell$ and $k = i+1$ to (4). Therefore we have proved that for all $j < k = i+1$ equation (9) holds.
Having shown that equation (1) is valid for $k = 2$ and if it holds for $k = i$ than it also holds for $k = i+1$ we can write that (1) holds for $k = 2, \ldots, K$ (by an inductive aplication) and it is just the condition of weak consistency.
(a) $\Rightarrow$ (b) If

$$\Theta = \{P_1((X_i)_{i \in C_1}), \ldots, P_K((X_i)_{i \in C_K})\}$$

is a weakly consistent system of distributions and further there is an ordering of C_k for $k = 1, \ldots, K$ meeting the *RIP*. Then through running the IPF procedure (see Definition 9) on this input set Θ we get a joint distribution $Q((X_i)_{i \in V}) \in \Pi(\Theta)$ just after K iterative steps and then the IPF procedure stops. Therefore for every input set having RIP and being weakly consistent there must exist a joint distribution from $\Pi(\Theta)$ and this is just the condition of strong consistency. $\square$

4 Where does inconsistency come from?

We believe that our world is consistent. A sentence and its negation can not be both together true. When a description of a part of this world is complete and correct it must be also consistent. We want to handle only correct information. But are all information we receive always complete? We are sure they are not. Should we ignore incomplete data? Let us use very simple example.

Example 1 (Problem of breathing difficulties) Suppose you are foreigner visiting the Czech Republic for the first time. During your visit you had talks with four czech people. You offered a cigarette to the first man you met at the Prague airport. But he declined saying he was from North Bohemia with very polluted air, he had enough smog there and that is why he need not to smoke. He also talked about his son who has breathing difficulties. Another man sitting beside of you said he was a smoker and he had no health difficulties, but you do not know where he was from. Later, when you were talking with your friend from small village in South Bohemia about relations between smoking, living in polluted area and breathing difficulties he mentioned himself as an example. He was not smoker. He lived in relatively clean area and he had not any breathing difficulties. The information you received is confirmed in a table (see Table 1) as data vectors.

Living in the North X_1	Smoker X_2	Breathing difficulties X_3
yes	no	?
yes	?	yes
?	yes	no
no	no	no

Table 1: Four Czechs answers

What could you conclude about problem of breathing difficulties from information you have been given? Most of them were not complete. In fact the only complete one was the information given by your friend from a village. Has it sense to integrate these incomplete information? In real life we are forced to use such information for our decisions very often. Our minds must be able to handle partial information. Artificial Intelligence, which has ambition to be in consideration in the least as good as people's one, should be able to cope with such problem. What are possible solutions? What are their advantages and shortages?

At first we might take only complete vectors. Than we can just sum and normalize for all combination over all complete vectors. In our example we receive distribution $Q_1(X_1, X_2, X_3)$ (in Table 3). Directly we can see it describes problem of breathing difficulties in very deformed way. In many situations of practice we have no complete data vectors, consequently this approach is unavailable.

Secondly we could replace each incomplete data vector u by n_u new vectors $c_1, \dots, c_{n_u}$. Each incomplete data vector u contains defined values of variables $(X_i)_{i \in V-U}$ while values of variables $(X_i)_{i \in U}$ are unknown. Each new vector $c_i, i = 1, \dots, n_u$ has the same values for known variables $(X_i)_{i \in V-U}$ as the original data vector u and one from all combinations of values of unknown variables $\times_{l \in U} \mathcal{X}_l$, hereat each combination is used just once. Therefore number of new complete vectors $c_i, i = 1, \dots, n_u$ is

$$n_u = \prod_{j \in U} \mid \mathcal{X}_j \mid .$$

We connect to each new vector u the weight $1/n_u$, all complete vectors would have weight 1. Afterwards we could calculate joint probability distribution just by weighted summing and normalizing over all combinations of variables' values. Distribution $Q_2(X_1, X_2, X_3)$ (see Table 3) was calculated this way. This distribution seems fine, but how long would our calculation take in a case of 100 variables with 90% missing values?

Lastly we have possibility to calculate several oligodimensinal distributions on subsets of variables from given problem domain and than integrate these pieces of information into one joint distribution (e.g. using IPF). These subsets may be put down by an expert or may be generated somehow automatically (see Section 9). Distribution $Q_3(X_1, X_2, X_3)$ was calculated by running IPF procedure on input set

$$\Theta_3 = \{P_1(X_1), P_2(X_2), P_3(X_3)\}.$$

Distributions $P_1 = (1/3, 2/3)$, $P_2 = (2/3, 1/3)$, $P_3 = (2/3, 1/3)$, were calculated from vectors which contained complete information on corresponding subsets $C_1 = \{X_1\}$, $C_2 = \{X_2\}$, $C_3 = \{X_3\}$.

Let us try to use as an input for IPF procedure set

$$\Theta_4 = \{P_1(X_1, X_2), P_2(X_1, X_3), P_3(X_2, X_3)\}.$$

Distribution calculated similarly as in previous case are in Table 2.

Table 2: Input distributions $\{P_1,\ P_2,\ P_3,\}$

P_1	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.5	0.0
$X_1 = 2$	0.5	0.0

P_2	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	0.5	0.0
$X_1 = 2$	0.0	0.5

P_3	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	0.5	0.0
$X_2 = 2$	0.5	0.0

It seems that problem of breathing difficulties is better described with input set Θ_4 than with Θ_3, because it is able to catch relations between the variables X_1, X_2 and X_3. But regular IPF procedure fails, because distributions in input set Θ_4 are *inconsistent.*

As we mentioned first approach can not be very often used because of frequent absence of complete vectors, second one would be unusable for higher dimension problems and third contains few information about given problem. In fact it can not involve any dependence. We would like to take as much information from data as it is possible (e.g. as it is in the fourth case), but then nobody can guarantee that input set of distribution would be consistent. The solution we offer is to modify IPF so that it converges to some reasonable joint distribution even in inconsistent case.

In Table 3 joint distributions and their entropy are compared. The distribution Q_4 is calculated from input set Θ_4 by modifications of IPF described in Sections 6 and 7.

Remark. From the point of view of mathematical statistics we need much more data vectors for making acceptable hypotheses about data. But in many practical situation this condition is out of reach. However last mentioned approach use the highest possible number of data vectors for estimation of oligodistributions.

Table 3: Comparison of joint distribution for problem of breathing difficulties

X_1	X_2	X_3	Q_1	Q_2	Q_3	Q_4
1	1	1	1	1/4	0.1481	4/9
1	1	2	0	0	0.0741	0
1	2	1	0	1/8	0.0741	2/9
1	2	2	0	0	0.0370	0
2	1	1	0	1/8	0.2963	0
2	1	2	0	1/4	0.1481	3/9
2	2	1	0	1/8	0.1481	0
2	2	2	0	1/8	0.0741	0
Entropy $H(Q_i)$			0	1.7329	1.9095	1.0609

5 Regular IPF in an inconsistent situation

Definition 9 (Iterative Proportional Fitting procedure) Let Q_0 be the joint uniform probability distribution defined for all the variables $(X_i)_{i\in V}$. Iterative Proportional Fitting procedure is the procedure which calculates joint distributions $Q_1, Q_2, \ldots$ by the formula

$$Q_n = Q_{n-1} \cdot \frac{P_j}{Q_{n-1}^{C_j}} \qquad \text{for} \quad j = ((n-1) \bmod K) + 1. \tag{10}$$

Theorem 2 (Convergence of IPF) *Sequence of the joint distributions calculated by the formula* 10 *is convergent iff* $\Theta = \{P_1, \ldots, P_k\}$ *is strongly consistent, i.e.* $\Pi(\Theta) \neq \emptyset$. *In the case of convergence*

$$Q = \lim_{n\to\infty} Q_n \in \Pi(\Theta),$$

and

$$H(Q) = \max_{R\in\Pi(\Theta)} H(R).$$

This theorem was proven in [Csis 75].

What can be said about the IPF process when the system Θ of oligodimensional distributions is not consistent, i.e. when $(\Pi(\Theta) = \emptyset)$?

After a number of computational experiments with the regular IPF procedure applied to both weakly consistent and inconsistent systems of oligodimensional distributions we believe that only two possibilities can happen.

Conjecture 1 *Let Q_0 be the joint uniform probability distribution defined for all the variables $(X_i)_{i\in V}$. For any system of strongly inconsistent oligodimensional distributions application of the IPF procedure (Definition 9) leads to one of the following two situations:*

- *The computation fails because there exits $1 < j \leq K$ such that the distribution P_j is not dominated by $Q_{j-1}^{C_j}$ (i.e. there exists an event x such that $P_j(x) > 0$ and $Q_{j-1}^{C_j}(x) = 0$), and thus the distribution Q_j is not defined.*
- *The sequence $Q_1, Q_2, \ldots$ tends to cycle. More precisely; for all $1 \leq j \leq K$ the subsequences $Q_j, Q_{j+K}, Q_{j+2K}, \ldots, Q_{j+nK}, \ldots$ are convergent. All the limit distributions for all these subsequences differ each from other.*

Let us illustrate these two possibilities on simple examples.

Example 2 (Three two-dimensional distributions) Consider three distributions of two variables each

$$\{P_1(X_1, X_2), P_2(X_2, X_3), P_3(X_1, X_3)\}$$

given in Table 4.

Table 4: Input distributions $\{P_1,\ P_2,\ P_3\}$

P_1	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	$0.5 - \varepsilon$	$0 + \varepsilon$
$X_1 = 2$	$0 + \varepsilon$	$0.5 - \varepsilon$

P_2	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	$0.5 - \varepsilon$	$0 + \varepsilon$
$X_2 = 2$	$0 + \varepsilon$	$0.5 - \varepsilon$

P_3	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	$0 + \varepsilon$	$0.5 - \varepsilon$
$X_1 = 2$	$0.5 - \varepsilon$	$0 + \varepsilon$

Taking $\varepsilon = 0$, the computation fails because the distribution P_3 is not dominated by $Q_2^{C_3}$. Namely, for $x_{12} = (X_1 = 1, X_3 = 2)$ $P_3(x_{12}) > 0$ and $Q_2^{C_3}(x_{12}) = 0$ as well as for $x_{21} = (X_1 = 2, X_3 = 1)$ $P_3(x_{21}) > 0$ and $Q_2^{C_3}(x_{21}) = 0$, and thus the distribution Q_3 is not defined. The respective distributions Q_1 and Q_2 are in Table 5.
Remark. In all examples mentioned in this paper we use the same test to stop iterative process when total variance of two sequential distribution $V(Q_{i-1}, Q_i)$ is less then 10^{-8}.

Table 5: Application of IPF to the distributions from Table 4 with $\varepsilon = 0$.

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3
1	1	1	0.1250	0.2500	0.5000	0.0000
1	1	2	0.1250	0.2500	0.0000	-
1	2	1	0.1250	0.0000	0.0000	0.0000
1	2	2	0.1250	0.0000	0.0000	-
2	1	1	0.1250	0.0000	0.0000	0.0000
2	1	2	0.1250	0.0000	0.0000	-
2	2	1	0.1250	0.2500	0.0000	0.0000
2	2	2	0.1250	0.2500	0.0000	-

The other situation occurs for $\varepsilon = 0.1$. In this case the sequence of distribution $Q_1, Q_2, \ldots$ "converges" to the cycle of the three distributions shown in Table 6. When the order of input set Θ is $\{P_3, P_2, P_1\}$ then the distributions in the cycle are different (see Table 7).

Remark. Let us stress that even for weakly consistent systems of oligodimensional distributions the systems of limit distributions may differ when the distributions in Θ are reordered.

6 Conservative modification of IPF

We are looking for a modification of IPF that meets the following **basic requirements**

Table 6: Application of IPF to the distributions from Table 4 with $\varepsilon = 0.1.$ and ordering $\{P_1, P_2, P_3\}$

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	...	Q_{19+3n}	Q_{20+3n}	Q_{21+3n}
1	1	1	0.1250	0.2000	0.3200	0.0941		0.1562	0.2439	0.1000
1	1	2	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
1	2	1	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
1	2	2	0.1250	0.0500	0.0800	0.2000		0.1000	0.1562	0.2438
2	1	1	0.1250	0.0500	0.0800	0.2000		0.1000	0.1562	0.2438
2	1	2	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
2	2	1	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
2	2	2	0.1250	0.2000	0.3200	0.0941		0.1562	0.2439	0.1000

Table 7: Application of IPF to the distributions from Table 4 with $\varepsilon = 0.1.$ and ordering $\{P_3, P_2, P_1\}$

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	...	Q_{19+3n}	Q_{20+3n}	Q_{21+3n}
1	1	1	0.1250	0.0500	0.0800	0.200		0.1000	0.1561	0.2438
1	1	2	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
1	2	1	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
1	2	2	0.1250	0.0200	0.3200	0.0941		0.1562	0.2439	0.1000
2	1	1	0.1250	0.0200	0.3200	0.0941		0.1562	0.2439	0.1000
2	1	2	0.1250	0.0500	0.0200	0.0059		0.0000	0.0000	0.0000
2	2	1	0.1250	0.2000	0.0800	0.2000		0.2438	0.1000	0.1562
2	2	2	0.1250	0.0500	0.0800	0.0200		0.1000	0.1561	0.2438

(i) if the input oligodimensional distributions are consistent, the procedure converges to the same distribution as the original IPF,

(ii) the procedure converges even in inconsistent situations,

(iii) the limit distribution is independent of the ordering of the input oligodimensional distributions,

(iv) and the marginals of the resulting distribution "are close" to the input oligodimensional distributions.

In this article two approaches are presented that seem to meet these requirements. At present we are testing their behavior and in this and next sections we summarize some of achieved experimental results.

First modification of IPF procedure was proposed by Radim Jiroušek and is described in [Jir,Vom 95]. It computes at each step of its process a convex mixture of the usually computed I-projection of Q_{i-1} into P_j and of the distribution corresponding to the previous step of iteration Q_{i-1}.

Definition 10 (Conservative modification of IPF) Let C_k for $k = 1, \ldots, K$ are subsets of index set V. For each C_k we assume that a probability distribution P_k defined for variables $(X_i)_{i \in C_k}$ is given. Further, let Q_0 be the joint uniform probability distribution defined for all the variables $(X_i)_{i \in V}$ and $\{\alpha_i\}$ is a monotonous sequence. Than the computational process defined by formula

$$Q_i = (1 - \alpha_i) \cdot Q_{i-1} + \alpha_i \cdot Q_{i-1} \cdot \frac{P_j}{Q_{i-1}^{C_j}} \qquad \text{for } j = ((i-1) \bmod K) + 1, \tag{11}$$

is entitled conservative modification of IPF.

Remark. Distribution $Q((X_i)_{i \in V})$ is defined for all combination of variables' values $x \in \times_{i \in V} \mathcal{X}_i$. Let $y \in \times_{i \in V-D} \mathcal{X}_i$ and $z \in \times_{i \in D} \mathcal{X}_i$, where $D \subset V$. Then symbol $Q(y \times z)$ is equivalent to $Q(x)$ and it is used to enable description of separate operation with $Q((X_i)_{i \in V})$ on a subset $\times_{j \in D} \mathcal{X}_j \subset \times_{i \in V} \mathcal{X}_i$.

Theorem 3 (Conservative modification convergence) *Let $\{\alpha_i\}$ is a monotonous sequence with $\lim_{i \to \infty} \alpha_i = 0$ and* for $i = 1, 2, \ldots \quad 0 \leq \alpha_i < 1$. *Than the sequence of joint distributions $Q_1, Q_2, \ldots$ defined by the formula* 11 *is convergent.*

Proof. By the definition of I-divergence and using formula (11) we can rewrite I-divergence of two sequential distribution:

$$\begin{aligned}
I(Q_{i-1} \parallel Q_i) &= \sum_{x \in \times_{l \in V} \mathcal{X}_l} Q_{i-1}(x) \cdot \log \frac{Q_{i-1}(x)}{Q_i(x)} \\
&= \sum_{x \in \times_{l \in V} \mathcal{X}_l} Q_{i-1}(x) \cdot \log \frac{Q_{i-1}(x)}{Q_{i-1}(x) \cdot (1 - \alpha_i + \alpha_i \cdot \frac{P_j(x)}{Q_{i-1}^{C_j}(x)})} \\
&= \sum_{x \in \times_{l \in V} \mathcal{X}_l} Q_{i-1}(x) \cdot \log \frac{1}{1 - \alpha_i + \alpha_i \cdot \frac{P_j(x)}{Q_{i-1}^{C_j}(x)}} \\
&= \sum_{y \in \times_{l \in C_j} \mathcal{X}_l} (\log \frac{1}{1 - \alpha_i + \alpha_i \cdot \frac{P_j(y)}{Q_{i-1}^{C_j}(y)}} \cdot \sum_{z \in \times_{m \in V - C_j} \mathcal{X}_m} Q_{i-1}(y \times z)) \\
&= \sum_{x \in \times_{l \in C_j} \mathcal{X}_l} (\log \frac{1}{1 - \alpha_i + \alpha_i \cdot \frac{P_j(x)}{Q_{i-1}^{C_j}(x)}} \cdot Q_{i-1}^{C_j}(x))
\end{aligned}$$

$$= \sum_{x \in \times_{l \in C_j} \mathcal{X}_l} Q_{i-1}^{C_j}(x) \cdot \log \frac{Q_{i-1}^{C_j}(x)}{\alpha_i \cdot P_j(x) + (1-\alpha_i) \cdot Q_{i-1}^{C_j}(x)}$$
$$= I(Q_{i-1}^{C_j} \parallel \alpha_i \cdot P_j(x) + (1-\alpha_i) \cdot Q_{i-1}^{C_j}(x)).$$

Then we can write:

$$\lim_{i \to \infty} I(Q_{i-1} \parallel Q_i) = \lim_{i \to \infty} I(Q_{i-1}^{C_j} \parallel \alpha_i \cdot P_j(x) + (1-\alpha_i) \cdot Q_{i-1}^{C_j}(x)) = I(Q_{i-1}^{C_j} \parallel Q_{i-1}^{C_j}) = 0.$$

Since $\sum_{x \in \times_{l \in V} \mathcal{X}_l} \mid P(x) - R(x) \mid \leq \sqrt{2 \cdot I(P \parallel R)}$, [Kull 67], then also

$$\lim_{i \to \infty} \sum_{x \in \times_{l \in V} \mathcal{X}_l} \mid Q_{i-1}(x) - Q_i(x) \mid = 0$$

and the sequence $Q_1, Q_2, \ldots$ converges. □

This modification has been tested with several types of input sets of oligodimensional distributions and with several different types of sequences of α_n as, for example,

$$\alpha_n = \frac{1}{1+n}, \qquad \alpha_n = \frac{1}{\sqrt{1+n}}, \qquad \alpha_n = \frac{1}{\pi}\left(\frac{\pi}{2} - arctg\ n\right).$$

The results achieved lead us to the following conjecture.

Conjecture 2 *If* $\sum_{n=1}^{\infty} \alpha_n = +\infty$ *then the conservative modification of IPF (formula* (11)*) meets the properties (i), (ii) and (iv) of the basic requirements. The limit distribution* Q *minimizes the sum of the total variances:*

$$\sum_{k=1}^{K} V(Q^{C_k}, P_k) = \min_R \sum_{k=1}^{K} V(R^{C_k}, P_k).$$

Remark. In contrast to the I-divergence, minimization of the sum of the total variances does not lead to a unique distribution but usually a whole convex set of distributions minimize the value $\sum_{k=1}^{K} |R^{C_k}(x) - P_k(x)|$ (see Table 12).

Example 3 (Another three two-dimensional distributions) Let us now consider another weakly consistent system of two-dimensional distributions

$$\{P_1(X_1, X_2),\ P_2(X_2, X_3),\ P_3(X_1, X_3)\}$$

given in Table 8. Results of application of the conservative modification of IPF to this system are shown in Tables 9 and 10. They contain results achieved for orderings P_1, P_2, P_3 and P_3, P_2, P_1 respectively. In both cases the sequence

$$\alpha_n = \frac{1}{1+i}, \qquad \text{for} \quad i = ((n-1) \text{ div } K) + 1$$

was used.

Table 8: Input distributions $\{P_1,\ P_2,\ P_3\}$

P_1	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.02	0.38
$X_1 = 2$	0.28	0.32

P_2	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	0.13	0.17
$X_2 = 2$	0.67	0.03

P_3	$X_3 = 1$	$X_3 = 2$
$X_1 = 1$	0.24	0.16
$X_1 = 2$	0.56	0.04

Table 9: Application of the conservative modification of IPF to the distributions from Table 8 with ordering $\{P_1, P_2, P_3\}$.

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	Q_4	$\cdots$	Q_{10001}
1	1	1	0.1250	0.0675	0.0557	0.0494	0.0358		0.0000
1	1	2	0.1250	0.0675	0.0624	0.0647	0.0469		0.0586
1	2	1	0.1250	0.1575	0.2546	0.2258	0.2411		0.2838
1	2	2	0.1250	0.1575	0.0866	0.0898	0.0959		0.0581
2	1	1	0.1250	0.1325	0.1093	0.1448	0.1584		0.1678
2	1	2	0.1250	0.1325	0.1226	0.0735	0.0804		0.0732
2	2	1	0.1250	0.1425	0.2304	0.3051	0.2958		0.3585
2	2	2	0.1250	0.1425	0.0784	0.0470	0.0456		0.0000

Table 10: Application of the conservative modification of IPF to the distributions from Table 8 with ordering $\{P_3, P_2, P_1\}$.

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	Q_4	$\cdots$	Q_{10001}
1	1	1	0.1250	0.1225	0.0858	0.0475	0.0450		0.0000
1	1	2	0.1250	0.1025	0.1010	0.0559	0.0599		0.0585
1	2	1	0.1250	0.1225	0.1875	0.2377	0.2251		0.2838
1	2	2	0.1250	0.1025	0.0600	0.0761	0.0815		0.0581
2	1	1	0.1250	0.2025	0.1417	0.1640	0.1759		0.1677
2	1	2	0.1250	0.0725	0.0715	0.0827	0.0641		0.0733
2	2	1	0.1250	0.2025	0.3100	0.2957	0.3172		0.3585
2	2	2	0.1250	0.0725	0.0425	0.0405	0.0314		0.0000

Table 11: Application of the method of iterative averages to the distributions from Table 8

X_1	X_2	X_3	Q_0	Q_1	Q_2	Q_3	...	Q_{41}
1	1	1	0.1250	0.0650	0.0345	0.0200		0.0000
1	1	2	0.1250	0.0583	0.0471	0.0464		0.0588
1	2	1	0.1250	0.2150	0.2507	0.2648		0.2839
1	2	2	0.1250	0.0950	0.0778	0.0704		0.0579
2	1	1	0.1250	0.1617	0.1648	0.1640		0.1680
2	1	2	0.1250	0.0817	0.0718	0.0718		0.0730
2	2	1	0.1250	0.2583	0.3219	0.3462		0.3584
2	2	2	0.1250	0.0650	0.0314	0.0164		0.0000

7 Method of iterative averages

Other modification was proposed by František Matúš [Mat 94]. It is based on the idea to prevent cycling by computing an average of all I-projections of Q_{i-1} into $\Pi(\{P_j\})$ for $k = 1, 2, \ldots, K$ at each step.

Definition 11 (Method of iterative averages) Let C_k for $k = 1, \ldots, K$ are subsets of index set V. For each C_k we assume that a probability distribution P_k defined for variables $(X_i)_{i \in C_k}$ is given. Further, let Q_0 be the joint uniform probability distribution defined for all the variables $(X_i)_{i \in V}$. Than the computational process defined by the formula

$$Q_n = \frac{1}{K} \sum_{j=1}^{K} Q_{n-1} \cdot \frac{P_j}{Q_{n-1}^{C_j}} \tag{12}$$

is entitled Method of iterative averages.

Conjecture 3 *The method of iterative averages (formula (12)) meets the properties (i), (ii), (iii), (iv) of the basic requirements. The limit distribution Q minimizes the sum of the total variances:*

$$\sum_{k=1}^{K} V(Q^{C_k}, P_k) = \min_R \sum_{k=1}^{K} V(R^{C_k}, P_k).$$

Remark. Limit distributions of the conservative modification and of the method of iterative averages need not to be identical (see Example 5).

Example 4 (Application of Method of Iterative Averages) Applying the method of iterative averages to the three two-dimensional distributions

$$\{P_1(X_1, X_2),\ P_2(X_2, X_3),\ P_3(X_1, X_3)\}$$

from Table 8 we get the limit distribution already at 41 steps. As it can be seen from Table 11, the limit distribution practically does not differ from the results achieved by the conservative modification of IPF.

Table 12: Comparison of the results achieved with the input system given in Table 8

X_1	X_2	X_3	Q^{a1}	Q^{a2}	Q^b	Q^c	Q^d
1	1	1	0.0000	0.0000	0.0000	0.0000	0.0000
1	1	2	0.0528	0.0578	0.0586	0.0588	0.0477
1	2	1	0.2874	0.2836	0.2838	0.2839	0.2898
1	2	2	0.0588	0.0577	0.0581	0.0579	0.0552
2	1	1	0.1682	0.1709	0.1678	0.1680	0.1695
2	1	2	0.0749	0.0703	0.0732	0.0730	0.0713
2	2	1	0.3579	0.3597	0.3585	0.3584	0.3665
2	2	2	0.0000	0.0000	0.0000	0.0000	0.0000
$\sum_{k=1}^{K} V(Q^{C_k}, P_k)$			0.440	0.440	0.440	0.440	0.440
$\sum_{k=1}^{K} I(Q^{C_k} \parallel P_k)$			0.0726	0.0736	0.0744	0.0744	0.0711

Table 13: Description of symbols used in Table 12

Q^{a1}	...	the barycentre distribution of the limit cycle, when regular IPF procedure is applied to the ordering $\{P_1, P_2, P_3\}$
Q^{a2}	...	the barycentre distribution of the limit cycle, when regular IPF procedure is applied to the ordering $\{P_3, P_2, P_1\}$
Q^b	...	the distribution computed with the conservative modification of IPF
Q^c	...	the distribution computed with the method of iterative averages
Q^d	...	the distribution minimizing the sum of Kullback-Leibler divergences from the input oligodimensional distributions

8 Comparisons of IPF modifications

We have presented two iterative methods converging even for inconsistent systems of oligodimensional distributions. Another distribution with similar properties can be received by application of the regular IPF procedure to the inconsistent system of distributions, and then, if the process starts to cycle, one can simply compute the average (barycentre) of the respective "limit" distributions. A disadvantage of this process is that it violates the condition (iii) of the basic requirements and that it cannot be sometimes used (see the first situation of Conjecture 1). Nevertheless, the results achieved in this way do not differ too much from those received by the conservative modification of IPF. An example supporting this statement is presented in Table 12 containing distributions computed for the system of distributions given in Table 8.

Example 5 (Four two-dimensional distributions) Comparison of results achieved by application of IPF procedure and its modifications on input set

$$\Theta_4 = \{P_1, P_2, P_3, P_4\}.$$

defined in Table 14 is in Table 15.

This example help us to avoid simple conclusions about behavior of the modifications. From table 15 we can see that even in consistent case ($\varepsilon = 0.05$) result distributions of both modifications differ however both modifications seem to converge to

Table 14: Input distributions $\{P_1, P_2, P_3, P_4\}$

P_1	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.02	0.38
$X_1 = 2$	0.28	0.32

P_2	$X_3 = 1$	$X_3 = 2$
$X_2 = 1$	0.13	0.17
$X_2 = 2$	0.67	0.03

P_3	$X_4 = 1$	$X_4 = 2$
$X_3 = 1$	0.1	0.7
$X_3 = 2$	0.15	0.05

P_4	$X_4 = 1$	$X_4 = 2$
$X_1 = 1$	$0 + \varepsilon$	$0.4 - \varepsilon$
$X_1 = 2$	$0.25 - \varepsilon$	$0.35 + \varepsilon$

Table 15: Comparison of results achieved with input set Θ_4 with $\varepsilon = 0.05$ (strong consistency) and $\varepsilon = 0.2$ (weak consistency)

X_1	X_2	X_3	X_4	$Q^a_{(\varepsilon=0.05)}$	$Q^b_{(\varepsilon=0.05)}$	$Q^c_{(\varepsilon=0.05)}$	$Q^a_{(\varepsilon=0.2)}$	$Q^b_{(\varepsilon=0.2)}$	$Q^c_{(\varepsilon=0.2)}$
1	1	1	1	0.0009	0.0009	0.0002	n.c.	0.0002	0.0000
1	1	1	2	0.0093	0.0094	0.0116	n.c.	0.0004	0.0000
1	1	2	1	0.0063	0.0062	0.0050	n.c.	0.0287	0.0285
1	1	2	2	0.0035	0.0035	0.0032	n.c.	0.0002	0.0000
1	2	1	1	0.0338	0.0336	0.0361	n.c.	0.1144	0.1155
1	2	1	2	0.3323	0.3325	0.3304	n.c.	0.2127	0.2124
1	2	2	1	0.0089	0.0089	0.0087	n.c.	0.0391	0.0394
1	2	2	2	0.0049	0.0050	0.0048	n.c.	0.0002	0.0000
2	1	1	1	0.0184	0.0186	0.0138	n.c.	0.0004	0.0000
2	1	1	2	0.1013	0.1013	0.1044	n.c.	0.1414	0.1416
2	1	2	1	0.1224	0.1221	0.1229	n.c.	0.0657	0.0669
2	1	2	2	0.0378	0.0379	0.0390	n.c.	0.0598	0.0597
2	2	1	1	0.0468	0.0472	0.0499	n.c.	0.0010	0.0000
2	2	1	2	0.2571	0.2565	0.2535	n.c.	0.3349	0.3539
2	2	2	1	0.0123	0.0125	0.0135	n.c.	0.0004	0.0000
2	2	2	2	0.0038	0.0039	0.0031	n.c.	0.0003	0.0000
Number of iterations				29	3157	292		10001	444
$\sum_{k=1}^{K} V(Q^{C_k}, P_k)$				0.000	0.004	0.000		0.212	0.200
$\sum_{k=1}^{K} I(Q^{C_k} \parallel P_k)$				0.00000	0.00001	0.00000		0.01315	0.01171

Table 16: Description of used symbols

$Q^a_{(\varepsilon=0.05)}$	...	the result distribution after a regular IPF procedure application for $\varepsilon = 0.05$,
$Q^b_{(\varepsilon=0.05)}$	...	the result distribution after an application of the conservative modification of IPF for $\varepsilon = 0.05$,
$Q^c_{(\varepsilon=0.05)}$	...	the result distribution after an application of the method of iterative averages for $\varepsilon = 0.05$,
$Q^a_{(\varepsilon=0.2)}$	...	the result distribution after a regular IPF procedure application for $\varepsilon = 0.2$,
$Q^b_{(\varepsilon=0.2)}$	...	the result distribution after an application of the conservative modification of IPF for $\varepsilon = 0.2$,
$Q^c_{(\varepsilon=0.2)}$	...	the result distribution after an application of the method of iterative averages for $\varepsilon = 0.2$,

the distributions with the same total variance $\sum_{k=1}^{K} V(Q^{C_k}, P_k)$. We can see different result in Table 12 where limit distributions of both modifications were the same. This different behavior of the modifications seems to be dependent on a shape of covering $(C_1, C_2, \ldots, C_K)$ of index set V which input distributions have.

9 Search for the best approximation of given data

In the next part of this paper we show to where could be narrowed search for the best approximation of given data by a decomposable model.

Definition 12 (Decomposable model) Probability distribution $P((X_i)_{i\in V})$ is called decomposable with respect to the system $\Theta = (D_1, D_2, \ldots, D_L)$ satisfying RIP if it factorizes with respect to Θ, which means that there exist real functions $\psi_1, \ldots, \psi_L$ defined on
$\times_{i\in D_1}\mathcal{X}_i, \ldots, \times_{i\in D_L}\mathcal{X}_i$ respectively, such that

$$P((X_i)_{i\in V}) = \prod_{l=1}^{L} \psi_l((X_i)_{i\in D_l})$$

In this section we will use instead of more exact $Q((X_i)_{i\in D_l} \mid (X_j)_{j\in D_m})$ symbol $Q((X_i)_{i\in D_l})^{D_l|D_m}$ which denotes conditional probability of $(X_i)_{i\in D_l}$ given $(X_j)_{j\in D_m}$. Because of very frequent occurrence of expression $D_l \cap \bigcup_{m=1}^{l-1} D_m$ we define

$$d(l) \stackrel{\text{def}}{=} D_l \cap \bigcup_{m=1}^{l-1} D_m = D_l \cap (D_1 \cup \ldots \cup D_{l-1}).$$

Let $P((X_i)_{i\in V})$ be theoretical distribution which we can access only through data vectors generated from $P((X_i)_{i\in V})$. We are looking for an an approximation of distribution P by a distribution $Q((X_i)_{i\in V})$ which could be rewritten

$$Q(X_i)_{i\in V} = \prod_{l=1}^{L} Q((X_j)_{j\in D_l})^{D_l|d(l)}, \tag{13}$$

where ordering of $\{D_1, D_2, \ldots D_L\}$ meets RIP. Equation (13) means that distribution $Q(X_i)_{i\in V}$ is decomposable with respect to the system $\Theta = (D_1, D_2, \ldots D_L)$. Distribution $Q((X_i)_{i\in V})$ has from all possible distributions defined on $\times_{i\in V}\mathcal{X}_i$ and factorizing on given covering $(D_1, D_2, \ldots, D_L)$ of set V strictly maximal entropy [Jir 92].

We want to minimize I-divergence between distribution $P((X_i)_{i\in V})$ and distribution $Q((X_i)_{i\in V})$ just by selecting the best covering $(D_1, D_2, \ldots, D_L)$ of set V.

$$\begin{aligned}
I(P \parallel Q) &= \sum_{x\in\times_{i\in V}\mathcal{X}_i} P(x)\cdot\log\frac{P(x)}{R(x)}\\
&= \sum_{x\in\times_{i\in V}\mathcal{X}_i} P(x)\cdot\log P(x) - \sum_{x\in\times_{i\in V}\mathcal{X}_i} P(x)\cdot\log Q(x)\\
&= -H(P) - \sum_{x\in\times_{i\in V}\mathcal{X}_i} P(x)\cdot\log\prod_{l=1}^{L} Q^{D_l|d(l)}(x)\\
&= -H(P) - \sum_{l=1}^{L}\sum_{x\in\times_{i\in V}\mathcal{X}_i} P(x)\cdot\log\frac{Q^{D_l}(x)}{Q^{d(l)}(x)}
\end{aligned}$$

Oligodistributions $Q^{D_1}, \ldots, Q^{D_L}$ are taken from data therefore $\forall l = 1, \ldots, L$ $P^{D_l} = Q^{D_l}$ and consequently $\forall l = 1, \ldots, L \quad P^{d(l)} = Q^{d(l)}$. In addition to this substitution we will do another one using definition of informational content (definition 5).

$$\begin{aligned}
&I(P \parallel Q) =\\
&= -H(P) - \sum_{l=1}^{L}\sum_{x\in\times_{i\in D_l}\mathcal{X}_i}(\log P^{D_l}(x) - \log P^{d(l)}(x))\cdot\sum_{y\in\times_{i\in V-D_l}\mathcal{X}_i} P(y\times x)\\
&= -H(P)\\
&\quad -\sum_{l=1}^{L}\Big(\sum_{x\in\times_{i\in D_l}\mathcal{X}_i} P^{D_l}(x)\cdot\log\frac{P^{D_l}(x)}{\prod\limits_{k\in D_l} P^{\{k\}}(x)} - \sum_{x\in\times_{i\in d(l)}\mathcal{X}_i} P^{d(l)}(x)\cdot\log\frac{P^{d(l)}(x)}{\prod\limits_{k\in d(l)} P^{\{k\}}(x)}\Big)\\
&\quad -\sum_{l=1}^{L}\Big(\sum_{x\in\times_{i\in D_l}\mathcal{X}_i} P^{D_l}(x)\cdot\log\prod_{k\in D_l} P^{\{k\}}(x) - \sum_{x\in\times_{i\in d(l)}\mathcal{X}_i} P^{d(l)}(x)\cdot\log\prod_{k\in d(l)} P^{\{k\}}(x)\Big)\\
&= -H(P) - \Big(\sum_{l=1}^{L} I(P^{D_l}) - \sum_{l=1}^{L} I(P^{d(l)})\Big) - F(P^{D_1}(x), \ldots, P^{D_L(x)}).
\end{aligned}$$

We used substitution

$$\begin{aligned}
&F(P^{D_1}(x), \ldots, P^{D_L}(x)) =\\
&= -\sum_{l=1}^{L}\Big(\sum_{x\in\times_{i\in D_l}\mathcal{X}_i} P^{D_l(x)}\cdot\sum_{k\in D_l}\log P^{\{k\}}(x) - \sum_{x\in\times_{i\in d(l)}\mathcal{X}_i} P^{d(l)}(x)\cdot\sum_{k\in d(l)}\log P^{\{k\}}(x)\Big)
\end{aligned}$$

and we rearrange this expression:

$$\begin{aligned}
&F(P^{D_1}(x), \ldots, P^{D_L}(x)) =\\
&= -\sum_{l=1}^{L}\Big(\sum_{k\in D_l}\sum_{x\in\mathcal{X}_k}\log P^{\{k\}}(x)\cdot\sum_{y\in\times_{i\in\{D_l-\{k\}\}}\mathcal{X}_i} P^{D_l}(x\times y)\\
&\qquad - \sum_{k\in d(l)}\sum_{x\in\mathcal{X}_k}\log P^{\{k\}}(x)\cdot\sum_{y\in\times_{i\in\{d(l)-\{k\}\}}\mathcal{X}_i} P^{d(l)}(x\times y)\Big)\\
&= -\sum_{l=1}^{L}\Big(\sum_{k\in D_l}\sum_{x\in\mathcal{X}_k}\log P^{\{k\}}(x)\cdot P^{\{k\}}(x) - \sum_{k\in d(l)}\sum_{x\in\mathcal{X}_k}\log P^{\{k\}}(x)\cdot P^{\{k\}}(x)\Big)\\
&= \sum_{l=1}^{L}\Big(\sum_{k\in D_l} H(P^{\{k\}}) - \sum_{k\in d(l)} H(P^{\{k\}})\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l=1}^{L} \sum_{k \in (D_l - d(l))} H(P^{\{k\}}) \\
&= \sum_{k \in V} H(P^{\{k\}}).
\end{aligned}
$$

we substitute back

$$F(P^{D_1}(x), \ldots, P^{D_L}(x)) = \sum_{k \in V} H(P^{\{k\}})$$

and than we receive

$$
\begin{aligned}
I(P \parallel Q) &= -H(P) && (14) \\
&\quad - \sum_{k \in V} H(P^{\{k\}} && (15) \\
&\quad - \sum_{l=1}^{L} (I(P^{D_l}) - I(P^{d(l)})). && (16)
\end{aligned}
$$

Expressions (14) and (15) depends only on given data and are independent of selected covering of the set V. Therefore we can select the best approximation of given data (in the sense of minimizing I-divergence $I(P \parallel Q)$) just by maximizing sum of differences of informational content of distribution on selected sets D_l and of informational content of the distribution on corresponding intersection $d(l)$.

10 Objective of next research

Main topics of next research follows from problems described in previous part of this paper. All of them can be classify to a domain usually called uncertainty in Artificial Intelligence. Our approach is almost based on results from probability theory and graph theory.

First area of our interest is integration methods of inconsistent knowledge. We want to do further tests of modification of Iterative Proportional Fitting procedure described in this paper. It is needed to give more precise description of their behavior. Results from the introductory part of our research were published in [Jir,Vom 95].

We have implemented Space-saving version of IPF. The program is described in [Kush 95]. There were several interesting problems we have touched during implementation phase. One of them is triangulation and we want to give attention to some triangulation algorithms together with tests of new designed one. A description of some of them can be found in [Jir 92, Kjær 90].

There are interesting models for handling uncertainty using conditional probabilities called Bayesian Nets within a frame of probabilistic models in AI. Introduction to their theory is in [Jen 93]. Bayesian Nets use decomposable model which give them availability to do computation of conditional probabilities fast enough. One from directions of research in the theory of Bayesian Nets is Learning of Bayesian Nets (e.g. [Heck 95, Lam,Bac 94]) and we are also interested in it.

References

[Csis 75] Csiszár, I., 1975, I-divergence geometry of probability distribution and minimization problems,*Ann.Prob.* 3: 146-158.

[Háj,Havr,Jir 92] Hájek, P., Havránek, T., and Jiroušek, R., 1992, "Uncertain Information Processing in Expert Systems," CRC Press, Inc., Boca Raton.

[Heck 95] Heckerman, D., 1995, A Tutorial on Learning Bayesian Networks (Technical Report), Microsoft Research Advanced Technology Division, Redmond.

[Jen 93] Jensen, F.,1993, "Introduction to Bayesian Networks", Aarlborg University, Aarlborg.

[Jir 92] Jiroušek, R., 1993, "Probabilistic Methods in Artificial Intelligence"(DrSc Thesis) Institute of Information Theory and Automation, Prague.

[Jir,Vom 95] Jiroušek, R., Vomlel, J., 1995, Inconsistent knowledge integration in a probabilistic model, *Proc. of Workshop "Mathematical Models for handling partial knowledge in A.I.,"* Plenum Publ. Corp., New York.

[Kell 64] Kellerer, 1964, Verteilungsfunktionen mit gegeben Marginalverteilungen, *Z. Warhsch. Verw. Gebeite*, 3: 126-127.

[Kjær 90] Kjærulff U., 1990, Triangulation of Graphs - Algorithm Giving Small Total State Space, Institute of Electronic Systems, Aarlborg.

[Kull 67] Kullback, 1967, A lower bound for discrimination information in terms of variation, *IEEE Trans. Inform, Theory* IT-13: 126-127.

[Kush 95] Kushmerick, N., 1995, Probabilistic Model Editor's (User Manual) Institute of Information Theory and Automation, Prague.

[Lam,Bac 94] Lam, W., Bacchus, F., 1994, Learning Bayesian Belief Networks: An approach based on the MDL Principle, *Computational Intelligence*, 10, 3, 269-293.

[Mat 94] Matúš, F., 1994, Personal communication.

[Per 83] Perez, A., 1983, A probabilistic approach to the integration of partial knowledge for medical decision-making (in Czech), *Proc. 1st Czechoslovak Congress of Biomed. Eng.*, Mariánské Lázně: 221-226.