Jiří Grim

On structural approximating multivariate discrete probability distributions

Kybernetika, Vol. 20 (1984), No. 1, 1--17

Persistent URL: http://dml.cz/dmlcz/125676

KYBERNETIKA — VOLUME 20 (1984), NUMBER 1

305

ON STRUCTURAL APPROXIMATING MULTIVARIATE DISCRETE PROBABILITY DISTRIBUTIONS

JIŘÍ GRIM

The purpose of structural approximating is to develop a flexible parametric model applicable to estimating various types of probability distributions. The approach combines the idea of product approximating with the concept of finite mixtures. To optimize mixtures of product approximations the maximum-likelihood principle is used. An efficient numerical solution is enabled by a convergent iterative procedure. A numerical example previously used by other authors is included to compare different structural approximations.

1. INTRODUCTION

Essentially, there are two possibilities how to estimate an unknown probability distribution if a sample of independent observations of a random vector is given. Applying a parametric method we assume the unknown distribution to be from a standard parametric family and estimate only the involved parameters. This approach is computationally feasible but the results are often unsatisfactory, since, in practical problems, the assumed model is usually not adequate. Nonparametric methods, on the other side, do not require special assumptions concerning the form of the estimated distribution. However, they are less efficient and have also some other disadvantages like exceeding storage requirements, poor properties in case of small sample size etc.

In the present paper we consider a special class of parametric probability distributions defined by a dependence structure of the involved random variables and by a set of marginal distributions. This "structural" approximating retains the computational simplicity of parametric methods but the form of the approximated distribution need not be a priori assumed. The structural approach is based on approximating multivariate probability distributions by products of their lower order marginals studied by several authors in different aspects.

Thus, to reduce storage requirements, Lewis [15] applied this approach to discrete probability distributions and suggested an information criterion, so called relative

entropy, as a measure of goodness of approximation. Introducing a notion of tree dependence Chow and Liu [3] considered approximating d-dimensional discrete distributions by $(d-1)$ second-order marginals under the same criterion. The optimal choice of marginal distributions, not studied by Lewis, is solved here in terms of graph theory — by finding the maximum-weight spanning tree. Chow and Liu extended their solution also to statistically formulated problems, when only estimated marginals are available. Similar situation is considered also in the papers [11] and [7] with the aim of approximating unknown multivariate distributions, especially when the sample size is small. Again the relative entropy is used as a criterion but, instead of dependence tree, a less general markovian dependence structure is optimized. The information theoretical context of the product approximating clarifies the paper of Perez [16] introducing general concepts of dependence tightness and ε-admissible simplification of dependence structures.

Several references on approximating discrete distributions relate to contingency tables. The problem introduced by Deming and Stephan [5], [18] is to estimate the cell probabilities $P_{ij}, (i = 1, 2, \ldots, r; j = 1, 2, \ldots, c)$ of an $r \times c$ contingency table for which the true marginal probabilities $p_{i.}, p_{.j}$ are known and fixed. Unfortunately, this is not the case when only a sample of observations is available. Nevertheless, considerable attention has been paid to this problem in the literature, especially to an interesting iterative procedure suggested by Deming and Stephan. To compute the estimates $\hat{p}_{ij}$ the procedure successively satisfies the corresponding marginal restrictions continuing this iteration until convergence. Extending the results of Lewis [15], Brown [1] suggested a similar procedure and showed that the relative entropy as a measure of goodness of approximation improves at each step of the iteration. Finally, emphasizing statistical aspects, Ireland and Kullback [10] thoroughly analysed the properties of this algorithm. Kullback [13] extended further its applicability to multivariate probability densities and with Ku [12] also to the sample-based estimating of discrete probability distributions. However, the algorithm of Deming and Stephan seems to be less favourable in higher dimensions, since the estimated distribution is iteratively computed at each point of the discrete space. Moreover, the underlying model involves a large number of parameters.

To develop structural approximations practically applicable to estimation problems we first recapitulate the concept of product approximating (Section 2). In Section 3 it is shown that, without any loss of generality, only a special type of dependence structure may be considered in estimation problems (Theorem 3.1). In Theorem 3.2 the original solution of Chow and Liu is presented in a rigorous form. The statistically oriented approximating based on maximum-likelihood estimates of parameters is considered in Section 4. A new type of approximations is developed by introducing finite mixtures. Finally (Section 5), a numerical example is used to illustrate the hierarchy of different structural approximations.

2. PRODUCT APPROXIMATIONS OF DISCRETE PROBABILITY DISTRIBUTIONS

Let us consider a probability distribution of a discrete random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$:

$$p(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathscr{X}; \quad \mathscr{X} = \mathscr{X}_1 \times \mathscr{X}_2 \times \ldots \times \mathscr{X}_d; \tag{2.1}$$

where $\mathscr{X}_k$, $(k = 1, 2, \ldots, d)$ is a set of discrete values of the variable X_k. To simplify notation of marginal distributions we denote by I the set of indexes $\{1, 2, \ldots, d\}$. If $A \subset I$ is a subset of I and $\bar{A} \subset I$ its complement:

$$I = \{1, 2, \ldots, d\}; \quad A = \{i_1, i_2, \ldots, i_k\} \subset I; \quad \bar{A} = \{i_{k+1}, \ldots, i_d\} = I \setminus A, \tag{2.2}$$

we denote by

$$\boldsymbol{x}_A = (x_{i_1}, \ldots, x_{i_k}) \in \mathscr{X}_A \equiv \mathscr{X}_{i_1} \times \ldots \times \mathscr{X}_{i_k}; \quad \boldsymbol{x}_{\bar{A}} = (x_{i_{k+1}}, \ldots, x_{i_d}) \in \mathscr{X}_{\bar{A}} \tag{2.3}$$

the corresponding subvectors and subspaces. Thus, to express a marginal distribution, we can write

$$p_A(\boldsymbol{x}_A) = p(\boldsymbol{x}_A) = p(x_{i_1}, x_{i_1}, \ldots, x_{i_k}) = \sum_{\boldsymbol{x}_{\bar{A}} \in \mathscr{X}_{\bar{A}}} p(\boldsymbol{x}); \quad \boldsymbol{x}_A \in \mathscr{X}_A; \quad A \subset I. \tag{2.4}$$

As indicated, the subscript A in p_A will be omitted whenever tolerable. (For $A = \emptyset$ we obtain $p(\boldsymbol{x}_\emptyset) = 1$.)

Now, if $\mathscr{A}$ is a partition of the set I:

$$\mathscr{A} = \{A_1, A_2, \ldots, A_K\}; \quad \bigcup_{k=1}^{K} A_k = I; \quad j \neq k \Rightarrow A_j \cap A_k = \emptyset \tag{2.5}$$

then, using conditional probability distributions, we can write the well known expansion formula

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_{A_K} \mid \boldsymbol{x}_{A_{K-1}}, \ldots, \boldsymbol{x}_{A_1})\, p(\boldsymbol{x}_{A_{K-1}} \mid \boldsymbol{x}_{A_{K-2}}, \ldots, \boldsymbol{x}_{A_1}) \ldots p(\boldsymbol{x}_{A_2} \mid \boldsymbol{x}_{A_1}). \tag{2.6}$$

The principle of product approximating naturally follows from equation (2.6) if one uses only subsets of the conditioning variables.

Definition 2.1. Let $\mathscr{A}$ be a partition of the set I and $\mathscr{S}$ a sequence of pairs of subsets with the property

$$\mathscr{S} = \{(A_1 \mid B_1), (A_2 \mid B_2), \ldots, (A_K \mid B_K)\}; \quad B_k \subset \bigcup_{j=1}^{k-1} A_j \subset I; \quad k = 1, \ldots, K; \quad (B_1 = \emptyset) \tag{2.7}$$

Further let $p(\boldsymbol{x}_A)$, $A \subset I$ be marginal distributions of a probability distribution p. Then the function

$$P(\boldsymbol{x} \mid \mathscr{S}) = \prod_{k=1}^{K} \frac{p(\boldsymbol{x}_{A_k}, \boldsymbol{x}_{B_k})}{p(\boldsymbol{x}_{B_k})} = p(\boldsymbol{x}_{A_1}) \prod_{k=2}^{K} p(\boldsymbol{x}_{A_k} \mid \boldsymbol{x}_{B_k}); \quad \boldsymbol{x} \in \mathscr{X} \tag{2.8}$$

will be called the product approximation of the distribution p and $\mathscr{S}$ is the dependence structure of this approximation.

Let us remark that the function (2.8) is itself a valid probability distribution. For the sake of simplicity we assume here and in the following that the involved conditional distributions exist or, respectively, the marginal distributions in the denominator are nonzero.

To measure the "closeness" of approximation we use a generally accepted information criterion – the relative entropy

$$\mathcal{H}(p, P) = \sum_{\boldsymbol{x}\in\mathscr{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{P(\boldsymbol{x} \mid \mathscr{S})} \geqq 0 \tag{2.9}$$

which is nonnegative and equals zero only if the two distributions p, P are identical. Expanding the logarithm in (2.9) and using the relation

$$\sum_{\boldsymbol{x}\in\mathscr{X}} - p(\boldsymbol{x}) \ln p(\boldsymbol{x}_A) = \sum_{\boldsymbol{x}_A\in\mathscr{X}_A} - p(\boldsymbol{x}_A) \ln p(\boldsymbol{x}_A) = H(p_A) = H(A)\,; \quad A \subset I \tag{2.10}$$

we can write

$$\mathcal{H}(p, P) = -H(I) + \sum_{k=1}^{K} [H(A_k \cup B_k) - H(B_k)] \tag{2.11}$$

where $H(A)$ denotes the Shannon entropy of the marginal distribution p_A. (For $A = \emptyset$ we define $H(p_\emptyset) = H(\emptyset) = 0$.) The relative entropy (2.11) may further be rewritten in the following equivalent forms

$$\begin{aligned} \mathcal{H}(p, P) &= -H(I) + \sum_{k=1}^{K} H(A_k \mid B_k) = \\ &= -H(I) + \sum_{k=1}^{K} H(A_k) - \sum_{k=1}^{K} \mathcal{I}(A_k, B_k) \end{aligned} \tag{2.12}$$

where $H(A \mid B)$ denotes the conditional Shannon entropy and $\mathcal{I}(A, B)$ is the mutual Shannon information between a pair of random vectors $\boldsymbol{X}_A$, $\boldsymbol{X}_B$:

$$H(A \mid B) = H(A \cup B) - H(B) = \sum_{\boldsymbol{x}_B\in\mathscr{X}_B} p(\boldsymbol{x}_B) \sum_{\boldsymbol{x}_A\in\mathscr{X}_A} - p(\boldsymbol{x}_A \mid \boldsymbol{x}_B) \ln p(\boldsymbol{x}_A \mid \boldsymbol{x}_B)\,; \tag{2.13}$$

$$\mathcal{I}(A, B) = H(A) + H(B) - H(A \cup B)\,; \quad A \cup B \subset I\,; \quad A \cup B = \emptyset\,.$$

Since the entropy $H(I)$ is constant for a fixed distribution p, we have the possibility to compare or optimize the product approximations (2.8) by maximizing the criterion

$$Q(\mathscr{S}) = -\sum_{k=1}^{K} H(A_k) + \sum_{k=1}^{K} \mathcal{I}(A_k, B_k) = -\sum_{k=1}^{K} H(A_k \mid B_k)\,, \tag{2.14}$$

i.e. independently of the actual probability distribution p. This important property, first noticed by Lewis [15], is employed also in Section 3 to derive a suitable general form of dependence structures.

From (2.14) it follows that, given a set of several marginal probability distributions, we can choose an optimal product approximation (2.8) by evaluating all alternatives

(cf. [15]). Another possibility to pose the approximation problem relates to the concept of ε-admissible simplification of dependence structures [16]. Assuming a known distribution p, we are interested in product approximations for which the approximation error $\mathcal{H}(p, P(.\mid \mathscr{S}))$, (cf. (2.9)) is less than a given positive ε. Obviously, the set of available marginal distributions may be complete or arbitrarily specified.

The statistically oriented formulation of the approximation problem is suitable when a direct estimation of the unknown multivariate distribution is difficult because of a limited sample size. In this case the estimates or marginal distributions up to a given order r may be expected to be more reliable. For this reason, developing the concept of structural approximating, we confine the general dependence structure (2.7) by the inequality

$$|A_k \cup B_k| \leqq r\,; \quad k = 1, 2, \ldots, K\,; \quad (1 \leqq r \leqq d)\,. \tag{2.15}$$

Nevertheless, despite this constraint the underlying problem of discrete optimization is extremely difficult.

Remark 2.1. One way to simplify the optimization problem is to set $B_k = A_{k-1}$, $(k = 2, 3, \ldots, K)$ in the sequence $\mathscr{S}$, (cf. (2.7)). We obtain a particular "markovian" dependence structure (cf. [11], [16], [7])

$$\mathscr{S} = \{(A_1 \mid \emptyset), (A_2 \mid A_1), \ldots, (A_K \mid A_{K-1})\}\,. \tag{2.16}$$

Unfortunately, the corresponding solution has the complexity of the travelling-salesman problem (cf. [7]) and is therefore intractable in higher dimensions.

3. OPTIMIZATION PROBLEM

First it will be shown, that optimizing the product approximations (2.8) under condition (2.15) we may confine ourselves only to a special class of dependence structures without any further loss of generality.

Theorem 3.1. Let $P(.\mid \mathscr{S})$ be a product approximation of a probability distribution p with the property (2.15). Then there is a product approximation $P(.\mid \mathscr{S}^*)$ with the dependence structure

$$\mathscr{S}^* = \{(A_1^* \mid B_1^*), \ldots, (A_K^* \mid B_K^*)\}\,; \quad |A_k^*| = 1\,; \quad |B_k^*| = \min\{r-1; k-1\}\,; \quad k = 1, 2, \ldots, d\,; \tag{3.1}$$

which is better in the sense of the inequality

$$Q(\mathscr{S}^*) \geqq Q(\mathscr{S}) \tag{3.2}$$

Proof. Let $\mathscr{S}'$ be a new dependence structure obtained by extending the sets B_k

in $\mathscr{S}$ consistently with the constraint (2.15):

$$\mathscr{S}' = \{(\mathsf{A}_1 \mid \mathsf{B}'_1), \ldots, (\mathsf{A}_K \mid \mathsf{B}'_K)\}\ ; \quad \mathsf{B}_k \subset \mathsf{B}'_k \subset \bigcup_{l=1}^{k-1} \mathsf{A}_l\ ; \tag{3.3}$$

$$|\mathsf{B}'_k| = \min\{r - |\mathsf{A}_k|;\ \sum_{l=1}^{k-1} |\mathsf{A}_l|\}\ .$$

Using formula (2.14) we obtain the inequality

$$Q(\mathscr{S}') - Q(\mathscr{S}) = \sum_{k=1}^{K} [H(\mathsf{A}_k \mid \mathsf{B}_k) - H(\mathsf{A}_k \mid \mathsf{B}'_k)] \geqq 0 \tag{3.4}$$

since each difference in the sum is nonnegative by the well known property of the Shannon entropy. Let us consider further a partition of a set A_l from $\mathscr{S}'$, ($|\mathsf{A}_l| > 1$):

$$\mathsf{A}_l = \mathsf{A}_l^{(1)} \cup \mathsf{A}_l^{(2)}\ ; \quad \mathsf{A}_l^{(1)} \cap \mathsf{A}_l^{(2)} = \emptyset\ ; \quad |\mathsf{A}_l^{(1)}| > 0\ ; \quad |\mathsf{A}_l^{(2)}| > 0\ . \tag{3.5}$$

If we modify the dependence structure $\mathscr{S}'$ according to the partition (3.5):

$$\mathscr{S}^+ = \{(\mathsf{A}_1 \mid \mathsf{B}'_1), (\mathsf{A}_2 \mid \mathsf{B}'_2), \ldots, (\mathsf{A}_l^{(1)} \mid \mathsf{B}_l^*), (\mathsf{A}_l^{(2)} \mid \mathsf{A}_l^{(1)} \cup \mathsf{B}'_l), \ldots, (\mathsf{A}'_K \mid \mathsf{B}'_K)\}\ ; \tag{3.6}$$

$$\mathsf{B}'_l \subset \mathsf{B}_l^* \subset \bigcup_{j=1}^{l-1} \mathsf{A}_j\ ; \quad |\mathsf{B}_l^*| = \min\{r - |\mathsf{A}_l^{(1)}|;\ \sum_{j=1}^{l-1} |\mathsf{A}_j|\}$$

then it holds

$$Q(\mathscr{S}^+) - Q(\mathscr{S}') = H(\mathsf{A}_l \mid \mathsf{B}'_l) - [H(\mathsf{A}_l^{(2)} \mid \mathsf{A}_l^{(1)} \cup \mathsf{B}'_l) + H(\mathsf{A}_l^{(1)} \mid \mathsf{B}_l^*)] = \tag{3.7}$$

$$= [H(\mathsf{A}_l \mid \mathsf{B}'_l) - H(\mathsf{A}_l^{(2)} \mid \mathsf{A}_l^{(1)} \cup \mathsf{B}'_l) - H(\mathsf{A}_l^{(1)} \mid \mathsf{B}'_l)] + [H(\mathsf{A}_l^{(1)} \mid \mathsf{B}'_l) - H(\mathsf{A}_l^{(1)} \mid \mathsf{B}_l^*)] =$$

$$= H(\mathsf{A}_l^{(1)} \mid \mathsf{B}'_l) - H(\mathsf{A}_l^{(1)} \mid \mathsf{B}_l^*) \geqq 0$$

since the last difference is again nonnegative. Consequently, if we partition all the sets A_l in $\mathscr{S}'$ having two or more elements (repeatedly, if necessary) and modify the dependence structure in a way suggested by (3.6), we obtain finally a dependence structure $\mathscr{S}^*$ with the ptoperty (3.1) satisfying the inequality:

$$Q(\mathscr{S}^*) - Q(\mathscr{S}') \geqq 0\ . \tag{3.8}$$

The proof is complete since the inequalities (3.4) and (3.8) imply the assertion (3.2). □

Theorem 3.1 simplifies the optimization problem considerably, since, without any loss of generality, the criterion $Q(\mathscr{S})$ may be maximized on the following reduced class of dependence structures, (cf. (2.7)):

$$\mathscr{T} = \{(i_1 \mid \mathsf{B}_1), (i_2 \mid \mathsf{B}_2), \ldots, (i_d \mid \mathsf{B}_d)\}\ ; \quad \pi = \pi(l) = (i_1, i_2, \ldots, i_d)\ ; \tag{3.9}$$

$$\mathsf{B}_k \subset \{i_1, i_2, \ldots, i_{k-1}\}\ ; \quad |\mathsf{B}_k| = \min\{r - 1;\ k - 1\}\ ; \quad k = 1, 2, \ldots, d\ ;$$

$$P(\mathbf{x} \mid \mathscr{T}) = \prod_{k=1}^{d} p(x_{i_k} \mid \mathbf{x}_{\mathsf{B}_k})\ ; \quad \mathbf{x} \in \mathscr{X}\ ;$$

where π is a permutation of the elements of the set l. Nevertheless, even in this form the optimization retains its complexity. Only for the case of two-dimensional marginals

$(r = 2)$ there is a surprisingly efficient solution of Chow and Liu [3] employing weighted graphs:

Theorem 3.2. Let $\mathcal{G}_d$ be a complete weighted graph over the set of vertices I:

$$\mathcal{G}_d = \{I, E \mid w\}\ ; \quad E = \{(j, k) : j \in I;\ k \in I;\ j \neq k\} \tag{3.10}$$

and the edge-weight function w be defined by equation

$$w : E \to \mathbb{R}_1\ ; \quad w(j, k) = I(j, k) = H(j) + H(k) - H(j, k)\ ; \tag{3.11}$$

Then the maximum weight spanning tree $\tau_d = \{I, F \mid w\}$; $F \subset E$ of the graph $\mathcal{G}_d$ defines a product approximation

$$P(\boldsymbol{x} \mid \sigma) = p(x_{i_1}) \prod_{k=2}^{d} p(x_{i_k} \mid x_{j_k}) = \left[\prod_{k=1}^{d} p(x_{i_k})\right] \left[\prod_{k=2}^{d} \frac{p(x_{i_k} x_{j_k})}{p(x_{i_k})\, p(x_{j_k})}\right] \tag{3.12}$$

with the dependence structure σ:

$$\sigma = \{(i_1 \mid -), (i_2 \mid j_2), \ldots, (i_d \mid j_d)\}\ ; \quad j_k \in \{i_1, i_2, \ldots, i_{k-1}\} \subset I\ ; \tag{3.13}$$

which maximizes the corresponding criterion $Q(\sigma)$, (cf. (2.14)):

$$Q(\sigma) = -\sum_{k=1}^{d} H(i_k) + \sum_{k=2}^{d} I(i_k, j_k) \tag{3.14}$$

Proof. Let us recall that an undirected graph is complete if every possible pair of vertices is joined by an edge. The degree of a vertex is defined as the number of edges incident on the vertex. A path between two vertices is a sequence of edges where each vertex in the sequence has degree two except for the initial and final vertex which have degree one. A graph is said to be connected if there exists a path between any pair of vertices in the graph. A circuit is a connected graph in which every vertex is of degree two. A tree is a connected graph with no circuits. A spanning tree of a graph is its subgraph which connects all vertices. Finally, the weight of a tree is the sum of the component edge weights.

To prove the theorem we show first, that any spanning tree $\tau_d = \{I, F \mid w\}$ defines a dependence structure of the type (3.13). Since the tree does not contain circuits, there is at least one vertex $i_d \in I$ with degree one in τ_d. Deleting this vertex and the corresponding edge $(i_d, j_d) \in F$; $j_d \in I \setminus \{i_d\}$ we obtain a subtree $\tau_{d-1} \subset \tau_d$:

$$\tau_{d-1} = \{I \setminus \{i_d\};\ F \setminus \{(i_d, j_d)\} \mid w\} \tag{3.15}$$

which also contains no circuits. Thus we can choose a vertex $i_{d-1} \in I \setminus \{i_d\}$ with degree one in τ_{d-1} and the corresponding edge $(i_{d-1}, j_{d-1}) \in F \setminus \{(i_d, j_d)\}$; $j_{d-1} \in I \setminus \setminus \{i_d, i_{d-1}\}$ Proceeding along this line we exhaust all $(d - 1)$ edges of the spanning tree τ_d and obtain finally an isolated vertex i_1. It is obvious, that the resulting sequences $i_1, i_2, \ldots, i_d; j_2, j_3, \ldots, j_d$; define a dependence structure σ, (cf. (3.13)) and thereby a product approximation (3.12). The weight of the spanning tree τ_d may be expressed

in the form

$$W(\tau_d) = \sum_{k=2}^{d} I(i_k, j_k) \,. \tag{3.16}$$

Let us assume now by contradiction that the product approximation $P(\,.\,|\,\sigma)$ does not maximize the criterion (3.14), i.e. that there is a dependence structure σ' satisfying the inequality

$$Q(\sigma') > Q(\sigma) \,. \tag{3.17}$$

It can be easily verified, that any dependence structure (3.13) uniquely defines a connected graph without circuits, i.e. σ' defines a spanning tree τ'_d of $\mathcal{G}_d$:

$$\tau'_d = \{I, F' \,|\, w\} \,; \quad F' = \{(i'_k, j'_k) : k = 2, 3, \ldots, d\} \tag{3.18}$$

Since τ_d has the maximum weight, we can write

$$W(\tau_d) = \sum_{k=2}^{d} I(i_k, j_k) \geqq W(\tau'_d) = \sum_{k=2}^{d} I(i'_k, j'_k) \,. \tag{3.19}$$

The proof is complete since, by inequality (3.17) we have:

$$\sum_{k=2}^{d} I(i'_k, j'_k) < \sum_{k=2}^{d} I(i_k, j_k) \,. \tag{3.20}$$

□

Thus, to apply Theorem 3.2, we have to evaluate the mutual information $I(i, j)$ for all pairs of random variables X_i, X_j

$$I(i, j) = \sum_{x_i \in \mathcal{X}_i} \sum_{x_j \in \mathcal{X}_j} p(x_i, x_j) \ln \frac{p(x_i, x_j)}{p(x_i)\, p(x_j)} \,; \quad i, j \in I \,; \quad i \neq j \tag{3.21}$$

and find the maximum-weight spanning tree of the corresponding complete graph. For this purpose a standard algorithm may be used (see e.g. [17]).

4. STRUCTURAL APPROXIMATING

In estimation problems only a sample of independent observations of a discrete random vector is available as a rule:

$$S = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\} \,; \quad \mathbf{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_d^{(n)}) \in \mathcal{X} \tag{4.1}$$

To construct product approximations in such situations we could replace the unknown marginal distributions by their respective sample estimates without formal difficulties. However, it is more justified to apply a parametric approach and the maximum-likelihood principle, even when (in the discrete case) the both approaches lead to the same optimization procedure (cf. [3]). In this sense, the product approximation (3.12) may be viewed as a parametric probability distribution defined by a dependence

structure σ and a set of two-dimensional marginals $\mathscr{P}_2$:

$$(4.2)\qquad P(\boldsymbol{x} \mid \sigma, \mathscr{P}_2) = p(x_{i_1}) \prod_{k=2}^{d} p(x_{i_k} \mid x_{j_k}) = p(x_{i_1}) \prod_{k=2}^{d} \frac{p(x_{i_k}, x_{j_k})}{p(x_{j_k})}\,; \quad \boldsymbol{x} \in \mathscr{X}\,;$$

$$\sigma = \{(i_1 \mid -), (i_2 \mid j_2), \ldots, (i_d \mid j_d)\}\,; \quad \mathscr{P}_2 = \{p_{ij};\, i, j \in I\}$$

To compute the maximum-likelihood estimates of these parameters we have to maximize the corresponding log-likelihood function

(4.3)

$$L(\sigma, \mathscr{P}_2) = \frac{1}{N} \sum_{n=1}^{N} \ln P(\boldsymbol{x}^{(n)} \mid \sigma, \mathscr{P}_2) = \frac{1}{N} \sum_{n=1}^{N} \ln p(x_{i_1}^{(n)}) + \sum_{k=2}^{d} \left[\frac{1}{N} \sum_{n=1}^{N} \ln p(x_{i_k}^{(n)} \mid x_{j_k}^{(n)}) \right]$$

Formula (4.3) may be rewritten in the form

(4.4)

$$L(\sigma, \mathscr{P}_2) = \sum_{\xi \in \mathscr{X}_{i_1}} \hat{p}_{i_1}(\xi) \ln p_{i_1}(\xi) + \sum_{k=2}^{d} \left[\sum_{\eta \in \mathscr{X}_{j_k}} \hat{p}_{j_k}(\eta) \sum_{\xi \in \mathscr{X}_{i_k}} \frac{\hat{p}_{i_k j_k}(\xi, \eta)}{\hat{p}_{j_k}(\eta)} \ln p_{i_k \mid j_k}(\xi \mid \eta) \right]$$

where

$$(4.5)\qquad \hat{p}_i(\xi) = \frac{1}{N} \sum_{n=1}^{N} \delta(\xi, x_i^{(n)})\,; \quad \hat{p}_{ij}(\xi, \eta) = \frac{1}{N} \sum_{n=1}^{N} \delta(\xi, x_i^{(n)})\, \delta(\eta, x_j^{(n)})\,;$$

$$i, j \in I\,; \quad \xi \in \mathscr{X}_i\,; \quad \eta \in \mathscr{X}_j\,; \quad \delta(\xi, \eta) = \begin{cases} 0; & \xi \neq \eta \\ 1; & \xi = \eta \end{cases}$$

are the sample estimates of the marginal probabilities. Consequently (cf. inequality (2.9)), for any fixed dependence structure σ, the likelihood function $L(\sigma, \mathscr{P}_2)$ is maximized by

$$(4.6)\qquad p_{i_1}(\xi) = \hat{p}_{i_1}(\xi)\,; \quad p_{i_k \mid j_k}(\xi \mid \eta) = \frac{\hat{p}_{i_k j_k}(\xi, \eta)}{\hat{p}_{j_k}(\eta)}\,; \quad \begin{array}{l} \xi \in \mathscr{X}_{i_k};\ \eta \in \mathscr{X}_{j_k}\,; \\ k = 2, 3, \ldots, d\,; \end{array}$$

After substitution (4.6) formula (4.4) may be rewritten in the form

$$(4.7)\quad L(\sigma, \hat{\mathscr{P}}_2) = \sum_{k=1}^{d} \sum_{\xi \in \mathscr{X}_k} \hat{p}_k(\xi) \ln \hat{p}_k(\xi) + \sum_{k=2}^{d} \sum_{\xi \in \mathscr{X}_{i_k}} \sum_{\eta \in \mathscr{X}_{j_k}} \hat{p}_{i_k j_k}(\xi, \eta) \ln \frac{\hat{p}_{i_k j_k}(\xi, \eta)}{\hat{p}_{i_k}(\xi)\, \hat{p}_{j_k}(\eta)}$$

and therefore, because of analogy between formulas (3.14) and (4.7), we may use Theorem 3.2 to compute the optimal dependence structure $\hat{\sigma}$. Formally, denoting

$$(4.8)\qquad \mathscr{I}_0(i, j) = \sum_{\xi \in \mathscr{X}_i} \sum_{\eta \in \mathscr{X}_j} \hat{p}_{ij}(\xi, \eta) \ln \frac{\hat{p}_{ij}(\xi, \eta)}{\hat{p}_i(\xi)\, \hat{p}_j(\eta)}\,; \quad i, j \in I$$

and using the maximum weight spanning tree, we obtain

$$(4.9)\qquad \hat{\sigma} = \arg\max_{\sigma} \left\{ \sum_{k=2}^{d} \mathscr{I}_0(i_k, j_k) \right\}\,;$$

(Possible ties in (4.9) may be decided arbitrarily.)

To improve the product approximation (4.2) we could consider marginal distributions of higher order ($r > 2$) in (3.9) but, unfortunately, solution of the corresponding optimization problem is not known. Another possibility provides the method of mixtures. Using product approximations as components of mixtures we obtain a new class of approximations with dependence structures of qualitatively higher complexity. To differentiate this class of distributions from the standard parametric types we use the term structural approximations. In this sense a product approximation may be viewed as a particular case of a structural approximation.

As the most simple example of a structural approximation let us consider first the "latent structure" model of Lazarsfeld (cf. [14], [6]):

$$(4.10) \qquad P(\mathbf{x} \mid \mathbf{W}, \mathscr{P}_1) = \sum_{m=1}^{M} w_m P(\mathbf{x} \mid \mathscr{P}_{1m}); \quad \mathbf{W} = (w_1, \ldots, w_M);$$

$$\mathscr{P}_1 = (\mathscr{P}_{11}, \ldots, \mathscr{P}_{1M});$$

$$P(\mathbf{x} \mid \mathscr{P}_{1m}) = \prod_{k=1}^{d} p(x_k \mid m); \quad \mathscr{P}_{1m} = \{p_i(. \mid m), i \in I\}; \quad \mathbf{x} \in \mathscr{X}.$$

Here the components are simple products of univariate conditional marginal distributions and a dependence structure is not explicitly involved as a parameter. The maximum-likelihood estimates of the parameters $\mathbf{W}$ and $\mathscr{P}_1$ may be computed by the following algorithm of Hasselblad-Shlezinger, (cf. [8]):

Step 1. Given the parameters $\mathbf{W}^{(t)}, \mathscr{P}_1^{(t)}, (t = 0, 1, \ldots)$ compute the quantities

$$(4.11) \quad p^{(t)}(m \mid \mathbf{x}^{(n)}) = \frac{w_m^{(t)} P(\mathbf{x}^{(n)} \mid \mathscr{P}_{1m}^{(t)})}{\sum_{j=1}^{M} w_j^{(t)} P(\mathbf{x}^{(n)} \mid \mathscr{P}_{1j}^{(t)})}; \quad m = 1, 2, \ldots, M; \quad n = 1, 2, \ldots, N;$$

Step 2. Compute the new parameters $\mathbf{W}^{(t+1)}, \mathscr{P}_1^{(t+1)}$ by Eqs.

$$(4.12) \quad w_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)}); \quad m = 1, 2, \ldots, M; \quad i = 1, 2, \ldots, d;$$

$$p_i^{(t+1)}(\xi \mid m) = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} \delta(\xi, x_i^{(n)}) p^{(t)}(m \mid \mathbf{x}^{(n)}); \quad \xi \in \mathscr{X}_i.$$

The convergence properties of this iterative procedure are discussed in [8], an extensive numerical example is described in [9].

The fundamental subject of this paper is the structural approximation obtained as a mixture of components (4.2):

$$(4.13) \qquad P(\mathbf{x} \mid \mathbf{W}, \boldsymbol{\sigma}, \mathscr{P}_2) = \sum_{m=1}^{M} w_m P(\mathbf{x} \mid \sigma_m, \mathscr{P}_{2m}); \quad \mathbf{x} \in \mathscr{X};$$

$$\mathbf{W} = (w_1, w_2, \ldots, w_M); \quad \boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_M); \quad \mathscr{P}_2 = (\mathscr{P}_{21}, \mathscr{P}_{22}, \ldots, \mathscr{P}_{2M})$$

$$P(\mathbf{x} \mid \sigma_m, \mathscr{P}_{2m}) = p(x_{i_1} \mid m) \prod_{k=2}^{d} \frac{p(x_{i_k}, x_{j_k} \mid m)}{p(x_{j_k} \mid m)};$$

$$\sigma_m = \{(i_1 \mid -), (i_2 \mid j_2), \ldots, (i_d \mid j_d)\}; \quad \mathscr{P}_{2m} = \{p_{ij}(., . \mid m); \quad i, j \in I\}$$

The underlying dependence structure σ of the probability distribution (4.13) is actually a mixture of dependence trees or briefly a "dependence forest". We show that, despite of its complexity, the structural approximation (4.13) may be efficiently optimized. To maximize the likelihood function

$$L(\mathbf{W}, \sigma, \mathscr{P}_2) = \frac{1}{N} \sum_{n=1}^{N} \ln \left[\sum_{m=1}^{M} w_m P(\mathbf{x}^{(n)} \mid \sigma_m, \mathscr{P}_{2m}) \right] \tag{4.14}$$

by a procedure analogous to the scheme (4.11), (4.12) we have to solve the following recurrent relations

(4.15)

$$(\sigma_m^{(t+1)}, \mathscr{P}_{2m}^{(t+1)}) = \arg\max_{\sigma_m, \mathscr{P}_{2m}} \left\{ \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)}) \ln P(\mathbf{x}^{(n)} \mid \sigma_m, \mathscr{P}_{2m}) \right\};$$

$$m = 1, 2, \ldots, M;$$

in a sufficiently simple explicit form (cf. [8]). Here the parenthesiszed sum is a weighted analogy of the likelihood function (4.3) and may be rewritten in the form (cf. (4.4)):

(4.16)

$$\mathscr{L}(\sigma_m, \mathscr{P}_{2m}) = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)}) \left[\ln p(x_{i_1}^{(n)} \mid m) + \sum_{k=2}^{d} \ln p(x_{i_k}^{(n)} \mid x_{j_k}^{(n)}, m) \right] =$$

$$= \sum_{\xi \in \mathscr{X}_{i_1}} p_{i_1}^{(t+1)}(\xi \mid m) \ln p_{i_1}(\xi \mid m) + \sum_{k=2}^{d} \sum_{\eta \in \mathscr{X}_{j_k}} \sum_{\xi \in \mathscr{X}_{i_k}} p_{i_k j_k}^{(t+1)}(\xi, \eta \mid m) \ln p_{i_k \mid j_k}(\xi \mid \eta, m);$$

where

$$p_i^{(t+1)}(\xi \mid m) = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} \delta(\xi, x_i^{(n)}) p^{(t)}(m \mid \mathbf{x}^{(n)}); \quad \xi \in \mathscr{X}_i; \quad \eta \in \mathscr{X}_j; \tag{4.17}$$

$$p_{ij}^{(t+1)}(\xi, \eta \mid m) = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} \delta(\xi, x_i^{(n)}) \delta(\eta, x_j^{(n)}) p^{(t)}(m \mid \mathbf{x}^{(n)}).$$

Similarly to Eq. (4.4), for any fixed dependence structure σ_m the weighted likelihood function (4.16) is maximized by

$$p_{i_1}(\xi \mid m) = p_{i_1}^{(t+1)}(\xi \mid M); \quad p_{i_k \mid j_k}(\xi \mid \eta, m) = \frac{p_{i_k j_k}^{(t+1)}(\xi, \eta \mid m)}{p_{j_k}^{(t+1)}(\eta \mid m)}; \tag{4.18}$$

$$k = 2, 3, \ldots, d; \quad \xi \in \mathscr{X}_{i_k}; \quad \eta \in \mathscr{X}_{j_k};$$

Again, having performed substitution (4.18) in Eq. (4.16), we obtain the formula

$$\mathscr{L}(\sigma_m, \mathscr{P}_{2m}^{(t+1)}) = \sum_{k=1}^{d} \sum_{\xi \in \mathscr{X}_k} p_k^{(t+1)}(\xi \mid m) \ln p_k^{(t+1)}(\xi \mid m) + \sum_{k=2}^{d} I^{(t+1)}(i_k, j_k \mid m) \tag{4.19}$$

where

$$I^{(t+1)}(i, j \mid m) = \sum_{\xi \in \mathscr{X}_i} \sum_{\eta \in \mathscr{X}_j} p_{ij}^{(t+1)}(\xi, \eta \mid m) \ln \frac{p_{ij}^{(t+1)}(\xi, \eta \mid m)}{p_i^{(t+1)}(\xi \mid m)\, p_j^{(t+1)}(\eta \mid m)} ; \tag{4.20}$$

$$i \neq j ; \quad i, j \in I ;$$

Consequently (cf. (4.7), (4.9)), we may compute the optimal dependence structure $\sigma_m^{(t+1)}$ by means of the maximum-weight spanning tree with the edge-weight function (4.20). Thus, the iterative procedure maximizing the likelihood function (4.14) may be summarized as follows:

Step 1. Given the parameters $\mathbf{W}^{(t)}, \boldsymbol{\sigma}^{(t)}, \mathscr{P}_2^{(t)}, (t = 0, 1, \ldots)$ compute the quantities

$$p^{(t)}(m \mid \mathbf{x}^{(n)}) = \frac{w_m^{(t)} P(x^{(n)} \mid \sigma_m^{(t)}, \mathscr{P}_{2m}^{(t)})}{\sum_{j=1}^{M} w_j^{(t)} P(\mathbf{x}^{(n)} \mid \sigma_j^{(t)}, \mathscr{P}_j^{(t)})} ; \quad \begin{matrix} m = 1, 2, \ldots, M ; \\ n = 1, 2, \ldots, N ; \end{matrix} \tag{4.21}$$

Step 2. Compute a new weight vector $\mathbf{W}^{(t+1)}$ by Eqs.

$$w_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)}) ; \quad m = 1, 2, \ldots, M ; \tag{4.22}$$

new estimates of the conditional marginal probabilities $\mathscr{P}_2^{(t+1)}$ by Eqs. (4.17) and a new dependence structure $\boldsymbol{\sigma}^{(t+1)}$, (cf. (4.20)) by Eqs.

$$\sigma_m^{(t+1)} = \arg\max_{\sigma_m} \{ \sum_{k=2}^{d} I^{(t+1)}(i_k, j_k \mid m) \} ; \quad m = 1, 2, \ldots, M ; \tag{4.23}$$

5. NUMERICAL EXAMPLE

An interesting comparison of structural approximations discussed in Section 4 suggests a numerical example used in [3] and [12]. A probability distribution $P^*(\mathbf{x})$, $\mathbf{x} = (x_1, x_2, x_3, x_4)$ of a vector of four binary variables defined by a table (cf. Table 1, column 2) is to be approximated. Thus, we are given a deterministic approximation problem but, formally, the maximum-likelihood procedures may also be used in this case if we construct an artificial sample S such that for each $\mathbf{x} \in \mathscr{X}$ the relative frequency of $\mathbf{x}$ in S is equal to $P^*(\mathbf{x})$.

First a simple product of four univariate marginals has been used as approximation

$$P_1(\mathbf{x}) = p_1(x_1)\, p_2(x_2)\, p_3(x_3)\, p_4(x_4) ; \tag{5.1}$$

$$p_1(1) = p_2(1) = p_3(1) = 0{\cdot}55 ; \quad p_4(1) = 0{\cdot}50 ; \quad (p_k(0) = 1 - p_k(1);)$$

Table 1. Comparison of structural approximations.

x_1	x_2	x_3	x_4	$P^*(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$P_5(x)$	$P_6(x)$	$P_7(x)$
0	0	0	0	·1000	·0456	·1296	·09977	·1037	·1000	·0857	·1000
0	0	0	1	·1000	·0456	·1037	·10000	·1296	·1000	·1143	·1000
0	0	1	0	·0500	·0557	·0370	·04958	·0296	·0500	·0600	·0500
0	0	1	1	·0500	·0557	·0296	·04927	·0370	·0500	·0400	·0500
0	1	0	0	·0000	·0557	·0152	·00051	·0149	·0000	·0000	·0000
0	1	0	1	·0000	·0557	·0121	·00026	·0124	·0000	·0000	·0000
0	1	1	0	·1000	·0681	·0681	·10011	·0669	·0833	·0900	·1000
0	1	1	1	·0500	·0681	·0546	·05035	·0558	·0667	·0600	·0500
1	0	0	0	·0500	·0557	·0530	·05027	·0519	·0600	·0643	·0500
1	0	0	1	·1000	·0557	·0636	·09996	·0648	·0900	·0857	·1000
1	0	1	0	·0000	·0681	·0152	·00039	·0148	·0000	·0000	·0000
1	0	1	1	·0000	·0681	·0182	·00076	·0185	·0000	·0000	·0000
1	1	0	0	·0500	·0681	·0331	·04945	·0397	·0400	·0500	·0500
1	1	0	1	·0500	·0681	·0397	·04978	·0331	·0600	·0500	·0500
1	1	1	0	·1500	·0832	·1488	·14992	·1785	·1667	·1500	·1500
1	1	1	1	·1500	·0832	·1785	·14962	·1488	·1333	·1500	·1500
Number of param.				15	4	7	28	9	14	14	15
Number of comp.				—	1	1	1	2	3	3	2
$\mathcal{H}(P^*, P_i)$				·0000	·3687	·0952	·0098	·0952	·0092	·0084	·0000

The distribution P_1 is given in the third column of Table 1, the achieved approximation error is expressed in terms of the relative entropy $\mathcal{H}(P^*, P_1)$ (the last row of Table 1). The quality of approximation considerably improves if one uses the tree-dependence product distribution (4.2). The corresponding solution of Chow and Liu having the form (cf. [3])

$$P_2(\mathbf{x}) = p_{4|1}(x_4 \mid x_1)\, p_{3|2}(x_3 \mid x_2)\, p_{2|1}(x_2 \mid x_1)\, p_1(x_1)\,; \tag{5.2}$$

$$\sigma = \{(1 \mid -), (2 \mid 1), (3 \mid 2), (4 \mid 1)\}\,;$$

$$p_1(1) = p_2(1) = p_3(1) = 0{\cdot}55\,; \quad p_4(1) = 0{\cdot}50\,;$$

$$p_{21}(1,1) = 0{\cdot}40\,; \quad p_{32}(1,1) = 0{\cdot}45\,; \quad p_{41}(1,1) = 0{\cdot}30\,;$$

is given in the fourth column. The number of independent parameters in P_2 is seven — except for the dependence structure σ. The fifth column contains an approximation suggested by Ku and Kullback [12]. Its accuracy is high but the underlying distribution (for $P_1(\mathbf{x})$ see (5.1)).

$$P_3(\mathbf{x}) = a(x_1, x_2)\, b(x_1, x_3)\, c(x_1, x_4)\, d(x_2, x_3)\, l(x_2, x_4) f(x_3, x_4)\, P_1(\mathbf{x}) \tag{5.3}$$

includes twenty eight independent parameters. Let us recall, that the approximation P_3 is numerically optimized at all points $\boldsymbol{x} \in \mathcal{X}$ and therefore formula (5.3) is actually not used.

The next three columns of Table 1 correspond to the mixtures of the form (4.10). The first one has two components and is characterized by nine independent parameters:

(5.4)

P_4 :

m	w_m	$p_1(1 \mid m)$	$p_2(1 \mid m)$	$p_3(1 \mid m)$	$p_4(1 \mid m)$
1	0·5500	0·2727	0·0000	0·1818	0·5455
2	0·4500	0·6667	1·0000	0·7778	0·4444

$p_i(0 \mid m) = 1 - p_i(1 \mid m)$;

The achieved quality of approximation is nearly the same as that of P_2. The distributions P_5, P_6 consist of three components:

(5.5)

P_5 :

m	w_m	$p_1(1 \mid m)$	$p_2(1 \mid m)$	$p_3(1 \mid m)$	$p_4(1 \mid m)$
1	0·4500	0·6667	1·0000	1·0000	0·4444
2	0·2500	1·0000	0·4000	0·0000	0·6000
3	0·3000	0·0000	0·0000	0·3333	0·5000

(5.6)

P_6 :

m	w_m	$p_1(1 \mid m)$	$p_2(1 \mid m)$	$p_3(1 \mid m)$	$p_4(1 \mid m)$
1	0·3500	0·4286	0·0000	0·0000	0·5714
2	0·4000	1·0000	1·0000	0·7500	0·5000
3	0·2500	0·0000	0·6000	1·0000	0·4000

The quality of these two approximations (each with fourteen independent parameters) is even better than that of P_3. Finally, structural approximation of the form (4.13) with two components has been used. The number of independent parameters is fifteen and therefore an exact reconstruction of the original distribution should be possible. This is actually obtained by the following distribution (cf. the last column):

(5.7) $P_7(\boldsymbol{x}) = w_1\, p_{4|2}(x_4 \mid x_2, 1)\, p_{3|2}(x_3 \mid x_2, 1)\, p_{2|1}(x_2 \mid x_1, 1)\, p_1(x_1 \mid 1) +$

$+ w_2\, p_{4|1}(x_4 \mid x_1, 2)\, p_{2|3}(x_2 \mid x_3, 2)\, p_{3|1}(x_3 \mid x_1, 2)\, p_1(x_1 \mid 2)$;

$w_1 = 0{\cdot}5000$; $\sigma_1 = \{(1 \mid -), (2 \mid 1), (3 \mid 2), (4 \mid 2)\}$;

$p_i(x_i \mid 1)$: $p_1(1 \mid 1) = p_2(1 \mid 1) = 0{\cdot}4000$; $p_3(1 \mid 1) = 0{\cdot}5000$; $p_4(1 \mid 1) = 0{\cdot}4333$;

$p_{ij}(x_i, x_j \mid 1)$: $p_{21}(1,1 \mid 1) = 0{\cdot}2000$; $p_{32}(1,1 \mid 1) = 0{\cdot}3000$; $p_{42}(1,1 \mid 1) = 0{\cdot}1333$;

$w_2 = 0{\cdot}5000$; $\sigma_2 = \{(1 \mid -), (3 \mid 1), (2 \mid 3), (4 \mid 1)\}$;

$p_i(x_i \mid 2)$: $p_1(1 \mid 2) = p_2(1 \mid 2) = 0{\cdot}7000$; $p_3(1 \mid 2) = 0{\cdot}6000$; $p_4(1 \mid 2) = 0{\cdot}5666$;

$p_{ij}(x_i, x_j \mid 2)$: $p_{31}(1,1 \mid 2) = 0{\cdot}3000$; $p_{23}(1,1 \mid 2) = 0{\cdot}6000$; $p_{41}(1,1 \mid 2) = 0{\cdot}4666$;

Remark 5.1. There is a topical question of structural approximating continuous probability distributions (cf. [2], [6], [11], [13]). It can be seen that, in continuous case, the most results of Sections 2 and 3 remain valid, though the evaluation of the integrals occuring in the criteria (2.9) or (2.14) may become uneasy. (This difficulty does not arise in estimation problems, if we use m.-l. method.)

Another problem relates to representing of marginal probability density functions. Whereas in a discrete case we have to estimate single probabilities, in continuous case the adequate nonparametric estimates are too awkward. One possibility to conserve computational efficiency of structural approximating is to apply a parametric approach at the level of marginal densities. In this way Gibson [6] generalized the "latent structure" – to the "latent profile" model by using products of univariate normal densities as components. Similarly, approximating the two-dimensional marginals by normal densities, we obtain a continuous analogy of the structural approximation (4.13):

$$(5.8)\qquad P(\mathbf{x} \mid \mathbf{W}, \boldsymbol{\sigma}, \mathcal{N}_2) = \sum_{m=1}^{M} w_m P(\mathbf{x} \mid \sigma_m, \mathcal{N}_{2m})\,; \quad \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}_d\,;$$

$$\mathcal{N}_2 = (\mathcal{N}_{21}, \mathcal{N}_{22}, \ldots, \mathcal{N}_{2M})\,; \quad \boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_M)\,;$$

$$P(\mathbf{x} \mid \sigma_m, \mathcal{N}_{2m}) = p(x_{i_1} \mid m) \prod_{k=2}^{d} p(x_{i_k} \mid x_{j_k}, m) = p(x_{i_1} \mid m) \prod_{k=2}^{d} \frac{p(x_{i_k}, x_{j_k} \mid m)}{p(x_{j_k} \mid m)}\,;$$

$$\sigma_m = \{(i_1 \mid -), (i_2 \mid j_2), \ldots, (i_d \mid j_d)\}\,; \quad \mathcal{N}_{2m} = \{p_{ij}(\cdot, \cdot \mid m)\,; \quad i, j \in I\}\,;$$

$$p_{ij}(x_i, x_j \mid m) = \frac{1}{\sqrt{((2\pi)^2 \det \mathbf{A}_{mij})}} \exp\{-\tfrac{1}{2}(x_i - c_{mi}, x_j - cm_j) \mathbf{A}_{mij}^{-1} (x_i - c_{mi}, x_j - c_{mj})^{\mathrm{T}}\}\,.$$

Here the resulting components $P(\mathbf{x} \mid \sigma_m, \mathcal{N}_{2m})$ are normal densities with specially sampled covariance matrices. To compute m.-l. estimates of the parameters $\mathbf{W}, \boldsymbol{\sigma}, \mathcal{N}_2$ we could derive the following iterative procedure (cf. [8])

Step 1.

$$(5.9)\quad p^{(t)}(m \mid \mathbf{x}^{(n)}) = \frac{w_m^{(t)} P(\mathbf{x})^{(n)} \mid \sigma_m^{(t)}, \mathcal{N}_{2m}^{(t)})}{\sum_{j=1}^{M} w_j^{(t)} P(\mathbf{x}^{(n)} \mid \sigma_j^{(t)}, \mathcal{N}_{2j}^{(t)})}\,; \quad \begin{matrix} m = 1, 2, \ldots, M\,; \\ n = 1, 2, \ldots, N\,; \end{matrix} \quad (t = 0, 1, \ldots)$$

Step 2.

$$(5.10)\quad w_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})\,; \quad c_{mi}^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} x_i^{(n)} p^{(t)}(m \mid \mathbf{x}^{(n)})\,;$$

$$a_{mij}^{(t+1)} = \frac{1}{\sum_{n=1}^{N} p^{(t)}(m \mid \mathbf{x}^{(n)})} \sum_{n=1}^{N} (x_i^{(n)} - c_{mi}^{(t+1)})(x_j^{(n)} - c_{mj}^{(t+1)}) p^{(t)}(m \mid \mathbf{x}^{(n)})\,; \quad i, j \in I\,;$$

$$\sigma_m^{(t+1)} = \arg\max_{\sigma_m} \left\{ \sum_{k=2}^{d} I^{(t+1)}(i_k, j_k \mid m) \right\}; \quad \mathbf{A}_{mij}^{(t+1)} = \begin{pmatrix} a_{mii}^{(t+1)} & a_{mij}^{(t+1)} \\ a_{mji}^{(t+1)} & a_{mjj}^{(t+1)} \end{pmatrix}$$

$$I^{(t+1)}(i, j \mid m) = -\ln\left[1 - \frac{(a_{mij}^{(t+1)})^2}{a_{mii}^{(t+1)}\, a_{mjj}^{(t+1)}}\right]; \quad i, j \in I; \quad i \neq j.$$

However, the problem includes actually estimating of conditional probability density functions and should be therefore analysed in more detail.

6. CONCLUSION

As mentioned in Section 4, the properties of structural approximations could be improved by including marginal distributions of higher order, i.e. by considering the components of the form (3.9) with $r > 2$. Using oriented graphs we could define the corresponding structures analogous to the dependence trees but their optimization is difficult because of some essential asymmetries. However, even in case of an acceptably simple solution one might prefer to use a structural approximation based on the second order marginals for computational reasons. Let us recall that optimizing the parameters of a mixture with M components of the form (3.9) we have to estimate $M \cdot \binom{d}{r}$ conditional marginal distributions at each iteration of the procedure (cf. (4.21)–(4.23)). From this point of view the structural approximating with dependence forest may appear as a reasonable compromise.

ACKNOWLEDGEMENT

Special thanks are due to doc. Dr. Ing. Jiří Beneš, DrSc., for his kind support during preparation of this paper.

(Received May 6, 1983.)

REFERENCES

[1] D. T. Brown: A note on approximations to discrete probability distributions. Inform. and Control *2* (1959), 4, 386–392.

[2] C. K. Chow: Tree dependence in normal distributions. International Symposium on Information Theory, Nordwijk, The Netherlands 1970.

[3] C. K. Chow and C. N. Liu: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inform. Theory *IT-14* (1968), 3, 462–467.

[4] C. K. Chow and T. J. Wagner: Consistency of an estimate of tree-dependent probability distribution. IEEE Trans. Inform. Theory *IT-19* (1973), 5, 369–371.

[5] W. E. Deming and F. F. Stephan: On a least squares adjustement of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. *11* (1940), 427–444.

[6] W. A. Gibson: Three multivariate models: factor analysis, latent structure analysis and latent profile analysis. Psychometrika *24* (1959), 229–252.

[7] J. Grim: On estimation of multivariate probability density functions for situation recognition in large scale systems. In: Proceedings of Third Formator Symposium (J. Beneš, L. Bakule, eds.), Academia, Prague 1979.

[8] J. Grim: On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions. Kybernetika *18* (1982), 3, 173—190.
[9] J. Grim: Application of finite mixtures to multivariate statistical pattern recognition. In: Proceedings of DIANA Conference held in Liblice near Prague, September 27 — October 1, 1982.
[10] C. T. Ireland and S. Kullback: Contingency tables with given marginals. Biometrika *55* (1968), 179—188.
[11] А. Д. Юдин: Об информативных структурах многомерных случайных величин. (On information structures of multidimensional random variables.) Известия АН СССР - Техническая кибернетика (1977), 6, 135—144.
[12] H. H. Ku and S. Kullback: Approximating discrete probability distributions. IEEE Trans. Inform. Theory *IT-15* (1969), 444—447.
[13] S. Kullback: Probability densities with given marginals. Ann. Math. Statist. *39* (1968), 4, 1236—1243.
[14] P. F. Lazarsfeld and N. W. Henry: Latent structure analysis. Houghton Mifflin, Boston 1968.
[15] P. M. Lewis: Approximating probability distributions to reduce storage requirements. Inform. and Control *2* (1959), 214—225.
[16] A. Perez: ε-admissible simplification of the dependence structure of a set random variables. Kybernetika *13* (1977), 6, 439—449.
[17] R. C. Prim: Shortest connection networks and some generalizations. Bell System Tech. J. *36* (1957), 1389—1401.
[18] F. F. Stephan: An iterative method of adjusting sample frequency tables when expected marginal totals are known. Ann. Math. Statist. *13* (1942), 166—178.

Ing. Jiří Grim, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8, Czechoslovakia.