KYBERNETIKA — VOLUME 34 (1998), NUMBER 1, PAGES 41 - 55

A METHOD FOR KNOWLEDGE INTEGRATION

MARTIN JANŽURA[1] AND PAVEL BOČEK[2]

With the aid of Markov Chain Monte Carlo methods we can sample even from complex multi-dimensional distributions which cannot be exactly calculated. Thus, an application to the problem of knowledge integration (e. g. in expert systems) is straightforward.

1. INTRODUCTION

The problem of knowledge integration is usually understood as a reconstruction of a joint global distribution from a collection of some marginals representing the knowledge base. In the case of highly dimensional distributions the computational aspect is crucial. A standard approach is based on assuming a suitable dependence structure (see Lauritzen [8] for a survey).

In the present paper we abandon some usually accepted assumptions, namely

(i) we do not expect consistency of the knowledge base;

(ii) we do not assume any decomposable form of the underlying global distribution.

Our approach consists of the following steps.

I. The marginal distributions (the knowledge base) are expressed in the Gibbsian form (as defined in Subsection 2.3) and converted (aggregated) into a collection of global distributions (Subsection 4.2).

II. The obtained global distributions are represented by a single global distribution, given as their barycenter (Subsection 5.2), which is understood as an estimate of the true underlying distribution.

III. The barycenter is numerically available up to its normalizing constant, which is, nevertheless, sufficient for sampling with some Markov Chain Monte Carlo algorithm.

IV. An empirical distribution from the sampled data is calculated as a final, numerically feasible estimate of the true distribution in Section 6. (Alternatively, we could find only the most likely configuration.)

The approach is illustrated with a "bootstrap-like" example in Section 7.

[1]Partially supported by the Grant Agency of the Czech Republic under Grant No. 202/96/0731.
[2]Partially supported by the Grant Agency of the Czech Republic under Grant No. 402/96/0414.

For the Gibbs distributions we can mention Besag [1] or Winkler [13] as basic references, and Moussouris [11] for the non-positive case. The concept of barycenter is adapted from Perez [12], see also Matúš [9]. For the Markov Chain Monte Carlo methods again Winkler [13] is a good reference, for more pratically oriented results see Gilks et al [5], while for this particular application see also Gelfand and Smith [3] or Janžura and Přeučil [7].

2. MODELING

2.1. State Space

Let us consider a product state space $X_S = \bigotimes_{s\in S} X_s$ of a discrete system. Here S is a finite set of indices (*sites, elements*), and X_s for each $s \in S$ is a finite state space.

For every $V \subset S$ we denote by $\mathrm{Pr}_V : X_S \to X_V = \bigotimes_{s\in V} X_s$ the corresponding projection function, and by $\mathcal{F}_V = \{f : X_V \to \overline{R}_-\}$ the set of all functions on X_V with values in the extended real line $\overline{R}_- = [-\infty, +\infty)$. For the sake of brevity we shall write $x_V = (x_s)_{s\in V}$ instead of $\mathrm{Pr}_V(x_S)$ for $x_S \in X_S$. We shall also not distinguish between $\mathcal{F}_V$ and $\mathcal{F}_V^* = \{f \circ \mathrm{Pr}_V; f \in \mathcal{F}_V\} \subset \mathcal{F}_S$.

For any finite set B we denote by $|B|$ its cardinality. In the present paper we assume the index set S to be rather large, i.e. $|S| \gg 1$. Therefore the system state space X_S will be extremely large, at least $|X_S| \geq 2^{|S|}$ since $|X_s| \geq 2$ for every $s \in S$. Such a number of possible states disables any probability measure on X_S from a direct general treatment. We cannot e.g. store all the values, we cannot sum over the set X_S, etc.

2.2. Markov Distributions

Let us consider a probability distribution P_S on the product state space X_S with the support $X_+(P_S) = \{x_S \in X_S; P_S(x_S) > 0\}$. As it was already emphasized in the preceding subsection, we cannot deal with the distribution P_S directly. Thus, we would like to define the distribution P_S through some simple treatable quantities.

We could make a substantial simplification assumption, namely, the *local Markov property* could be assumed, i.e.

$$P_{s|S\setminus\{s\}}\left(x_s|x_{S\setminus\{s\}}\right) = P_{s|\partial s}(x_s|x_{\partial s}) \quad \text{for } x_S \in X_+(P_S)$$

where $\partial s \subset S \setminus \{s\}$ is a neighborhood of s for each $s \in S$. The *neighborhood system* $\{\partial s\}_{s\in S}$ obeys the symmetry property: $t \in \partial s$ iff $s \in \partial t$. Obviously we have $P_{t|\partial t} P_{s|S\setminus\{s,t\}} = P_{s|\partial s} P_{t|S\setminus\{s,t\}}$ on $X_+(P_S)$. Therefore, if $\partial s \subset S \setminus \{s,t\}$ then $\partial t \subset S \setminus \{s,t\}$ and vice versa. A graph $\mathcal{G}$ is induced on the set S by the system of neighboring pairs: $\langle s,t\rangle \in \mathcal{G}$ iff $s \in \partial t$.

In practical applications the Markov assumption is usually fully justified thanks to the physical experience of local interactions in the nature. Unfortunately, a complex system of constrains has to be satisfied in order to obtain a consistent collection of *local characteristics* $\{P_{s|\partial s}\}_{s\in S}$. Therefore, as we shall see in the following subsection, the Gibbsian approach is much more convenient.

2.3. Gibbs Distributions

Every probability distribution can be expressed in the *Gibbsian* form, i.e.

$$P_S(x_S) = P_S^{\Phi}(x_S) = \frac{1}{Z_S^{\Phi}} \exp\left\{\sum_{A\in\mathcal{A}} \Phi_A(x_A)\right\}, \qquad \text{(GD)}$$

where the system $\Phi = \{\Phi_A\}_{A\in\mathcal{A}}$ is called a *potential*, the particular functions $\Phi_A \in \mathcal{F}_A$ are quoted as *interactions*, $\mathcal{A} \subset \exp S\backslash\{\emptyset\}$, and $Z_S^{\Phi} = \sum_{x_S\in X_S} \exp\{\sum_{A\in\mathcal{A}} \Phi_A(x_A)\}$ is the normalizing constant. We assume there exists some $x_S^0 \in X_S$ with

$$\sum_{A\in\mathcal{A}} \Phi_A(x_A^0) > -\infty,$$

and we accept the standard conventions for calculating, i.e. $e^{-\infty} = 0$, $\log 0 = -\infty$, $-\infty + c = -\infty$ for $c \in \overline{R}_-$, etc.

For the positive distribution P_S (which is the much more easier case) the interactions can be obtained e.g. by the *Möbius formula*

$$\Phi_V(x_V) = \sum_{B\subset V} (-1)^{|V\setminus B|} \log P_S(x_B, 0_{S\setminus B}) \qquad \text{(MF)}$$

for each $x_V \in X_V$ and $V \in \mathcal{A}$, where $0_S = (0_s)_{s\in S} \in X_S$ is some fixed basic configuration ("vacuum", as it is called in the frame of statistical physics).

Then, from the definition (MF) we can directly observe

i) $\Phi_V(x_V) = 0$ if $x_s = 0_s$ for some $s \in V$.

ii) Moreover, if P_S is a Markov distribution, we have $\Phi_V \equiv 0$ if V is not a clique in the graph $\mathcal{G}$.

Under the normalizing condition i) there is a one-to-one relation (given by (GD) and (MF)) between positive probability distributions and potentials. Due to the latter property ii) we may always set $\mathcal{A} = \mathcal{C}$, where $\mathcal{C}$ is the system of all cliques in $\mathcal{G}$ (including one-body sets).

Since each P_S^{Φ} is obviously a Markov distribution with the neighborhood system

$$\left\{\partial s = \bigcup_{A\in\mathcal{A}, s\in A} A \setminus \{s\}\right\}_{s\in S},$$

the equivalence between Gibbs and Markov distributions is established.

In the general case, i.e. without the positivity assumption, the treatment is more complicated (see [11]). Anyhow, we can always set

$$\Phi_S(x_S) = \log P_S(x_S)$$

for $x_S \in X_S$ (with $\Phi_S(x_S) = -\infty$ for $x_S \notin X_+(P_S)$), and $\Phi_V \equiv 0$ otherwise.

The main advantage of the Gibbsian approach consists in an absence of any additional condition on the potential to compare with the system of local characteristics. Thus, we may start directly with a potential

$$\Phi = \{\Phi_A \in \mathcal{F}_A\}_{A\in\mathcal{A}}.$$

Let us emphasize again that in most cases the normalizing constant Z_S^Φ of a Gibbs distribution is numerically not available. The problem is inherent and cannot be easily avoided. Therefore the Monte Carlo methods are so important.

3. KNOWLEDGE INTEGRATION

3.1. Knowledge Base

The knowledge base will be given as a collection of probability distributions

$$\widetilde{Q}^j_{B_j}, \quad j = 1, \ldots, K,$$

each $\widetilde{Q}^j_{B_j}$ being defined on the product space $X_{B_j} = \bigotimes_{s \in B_j} X_s$ where $B_j \subset S$ is "reasonably small" for every $j = 1, \ldots, K$. For the sake of simplicity we assume the collection of subsets $\mathcal{B} = \{B_j\}_{j=1,\ldots,K}$ to create a covering of S, i.e. $\bigcup_{j=1}^K B_j = S$.

The probability distributions $\widetilde{Q}^j_{B_j}$, $j = 1, \ldots, K$, are supposed to be given. Let us emphasize that

i) we do not assume these base distributions to be mutually consistent,

ii) nor do we assume them to be consistent with the "true" underlying distribution P_S.

Remark. The inconsistencies can be caused, as we may imagine, by various sources of information and various kinds of errors (e.g. we deal with empirical distributions extracted from different data sets). These input errors can naturally imply also some inaccuracies in the outcomes. Nevertheless, and this is the definite essence of the above assumption, the proposed method can be applied without checking the consistency and irrespectively of its absence.

3.2. Problem and Solution

A knowledge base $\widetilde{Q}^j_{B_j}$, $j = 1, \ldots, K$ being given, we are interested in some estimates $\widehat{P}_V$ or $\widehat{P}_{V|W}$ of the marginal P_V, $V \subset S$, or the conditional marginal $P_{V|W}$, $V \subset S$, $W \subset S \setminus V$, respectively, in case of reasonable small $V \subset S$.

Our solution consists of four steps:

I. *Aggregation.* We convert the knowledge base $\{\widetilde{Q}^j_{B_j}\}_{i=1,\ldots,K}$ into a collection of distributions $\boldsymbol{Q} = \{Q^i_S\}_{i=1,\ldots,M}$ with each Q^i_S defined on X_S.

II. *Representation.* We choose a single distribution Q^0_S representing the collection $\boldsymbol{Q}$.

III. *Simulation.* For small enough $V \subset S$ we simulate a sequence of configurations

$$x_S^{(1)}, \ldots, x_S^{(n)}$$

sampled from Q^0_S, resp.

$$x^{(1)}_{S\setminus W}, \ldots, x^{(n)}_{S\setminus W}$$

sampled from $Q^0_{S\setminus W|W}(\cdot|\overline{x}_W)$ with some fixed $\overline{x}_W \in X_W$. We apply some of the MCMC algorithms described in Section 6. This is the crucial point of the method: In spite of not being able to calculate the distribution Q^0_S, we can perform the sampling.

IV. *Calculation.* Finally, based on these data, we calculate the empirical distribution

$$\{\widehat{P}_V(y_V)\}_{y_V \in X_V} \quad \text{where} \quad \widehat{P}_V(y_V) = \frac{1}{n}\sum_{r=1}^{n} \delta(y_V, x^{(r)}_V)$$

resp.

$$\{\widehat{P}_{V|W}(y_V|\overline{x}_W)\}_{y_V \in X_V} \quad \text{where} \quad \widehat{P}_{V|W}(y_V|\overline{x}_W) = \frac{1}{n}\sum_{r=1}^{n} \delta(y_V, x^{(r)}_V).$$

The empirical distributions are understood as estimates of the true quantities.

Remark. Let us recall that in spite of sampling from a multidimensional distribution Q^0_S, we actually calculate only $\widehat{P}_V$ with rather small $V \subset S$ which correspond to the variables under interest. Thus the empirical distribution $\widehat{P}_V$ can converge fast enough.

For $V \subset S$ large, when calculating and storing all $\widehat{P}_V(x_V)$ resp. $\widehat{P}_{V|W}(x_V|x_W)$ for every $x_V \in X_V$ would be hardly possible, we still might be interested in a highly likely configuration $x_V \in X_V$ (cf. Janžura and Přeučil [7]).

With the aid of the simulated annealing algorithm (cf. Subsection 6.4 below) we can find

$$\widehat{x}_S \in \operatorname*{arg\,max}_{x_S \in X_S} Q^0_S(x_S)$$

resp.

$$\widehat{x}_{S\setminus W} \in \operatorname*{arg\,max}_{x_{S\setminus W} \in X_{S\setminus W}} Q^0_{S\setminus W|W}(x_{S\setminus W}|\overline{x}_W).$$

The projection $\widehat{x}_V$ can be understood as a solution now. Such solution agrees with the Bayesian inference Using Gibbs Sampling (BUGS – cf. Gilks et al [5]).

3.3. Errors

There are several kinds of error in the proposed method.

i) First, some error might and usually will be in the knowledge base. But the reliability of the input information is not a subject of the present paper.

ii) Another error (or a loss of information) can be caused by the aggregation and the representation steps. Here we have to proceed carefully in order to preserve as much information as possible. See Sections 4 and 5.

The above errors can be understood as the "approximation errors" and can be expressed as a distinction between the true distribution P_S and the representing Q^0_S.

iii) In the simulation step a different type, namely the statistical error occurs. It can be described e. g. by the variance of the empirical distribution $\widehat{P}_V$ (resp. $\widehat{P}_{V|W}$) and can be decreased by a growing sample size.

4. AGGREGATION

4.1. Aggregation principle

We would like to represent the knowledge base by a single distribution Q_S^0 which should be an approximation of the underlying distribution P_S. Unfortunately, we cannot deal directly with the distributions $\widetilde{Q}_{B_j}^j$ since any iterative calculation (cf. e. g. Matúš [9]) is unfeasible again due to the dimensionality. Therefore, we convert at first the knowledge base into a collection of distributions on X_S, i. e.

$$\{\widetilde{Q}_{B_j}^j\}_{j=1,\dots,K} \longrightarrow \{Q_S^i\}_{i=1,\dots,M} = \boldsymbol{Q}.$$

Since we shall usually deal with rather small sets B_j we believe to obtain M much smaller than K and therefore we call this step as the *knowledge aggregation*.

There is no unique way of such a procedure. We should insist on a preservation of the information, i. e. on some kind of correspondence between the knowledge base and the collection $\boldsymbol{Q}$.

Thus, we introduce the aggregation principle (AP):

> With a minimal M, every $\widetilde{Q}_{B_j}^j$ should be the corresponding marginal of some Q_S^i, i. e.
>
> $$\widetilde{Q}_{B_j}^j = Q_{B_j}^i$$
>
> for some $i \in \{1, \dots, M\}$. Or, in other words, every $\widetilde{Q}_{B_j}^j$ should find its extension Q_S^i in $\boldsymbol{Q}$.

Let us consider the following procedure. By $\mathcal{S} = \{\sigma : \{1, \dots, K\} \to \{\sigma(1), \dots, \sigma(K)\}\}$ we denote the set of all permutations. For $\sigma \in \mathcal{S}$ we set

$$Q_S^\sigma = \prod_{j=1}^{M^\sigma} \widetilde{Q}_{B_{\sigma(j)}^{\sigma,1}|B_{\sigma(j)}^{\sigma,2}}^{\sigma(j)}$$

where

$$B_{\sigma(j)}^{\sigma,1} = B_{\sigma(j)} \setminus \bigcup_{\ell=1}^{j-1} B_{\sigma(\ell)}, \quad B_{\sigma(j)}^{\sigma,2} = B_{\sigma(j)} \cap \bigcup_{\ell=1}^{j-1} B_{\sigma(\ell)}$$

and M^σ is naturally given by $\bigcup_{j=1}^{M^\sigma} B_{\sigma(j)} = S$.

Let us emphasize that the probability distributions Q_S^σ can be expressed in the Gibbsian form (see Subsection 2.3). We may set $\mathcal{A} = \{B_{\sigma(j)}; B_{\sigma(j)}^{\sigma,1} \neq \emptyset\}$ and

$$\Phi_{B_{\sigma(j)}}^\sigma(x_{B_{\sigma(j)}}) = \log \widetilde{Q}_{B_{\sigma(j)}^{\sigma,1}|B_{\sigma(j)}^{\sigma,2}}(x_{B_{\sigma(j)}^{\sigma,1}}|x_{B_{\sigma(j)}^{\sigma,2}})$$

for $B_{\sigma(j)} \in \mathcal{A}$. (Thanks to the properties of conditional distributions the definition is correct.)

For a collection of permutations $\sigma^1, \ldots, \sigma^M$ we obtain a collection of distributions $\{Q_S^j = Q_S^{\sigma^j}\}_{j=1,\ldots,M}$. If $B_{\sigma(1)}, \ldots, B_{\sigma(L)}$ are pair-wise disjoint then Q_S^σ is an extension of every $\widetilde{Q}_{B_{\sigma(\ell)}}^{\sigma(\ell)}$ for $\ell = 1, \ldots, L$.

4.2. A method of aggregation

Based on the above observation we can propose a the following procedure:

i) Find a minimal covering of the system $\mathcal{B}$, namely $\mathcal{B} = \bigcup_{j=1}^M \mathcal{B}^j$ where each $\mathcal{B}^j = \{B_{j_1}, \ldots, B_{j_{m_j}}\} \subset \mathcal{B}$ consists of pairwise disjoint sets.

For an algorithm see Subsection 4.3 below.

ii) For $j = 1, \ldots, M$ set

$$\sigma^j(\ell) = j_\ell \qquad \text{for} \quad \ell = 1, \ldots, m_j$$

$$\sigma^j(\ell) \in \{1, \ldots, M\} \setminus \{\sigma^j(k); k = 1, \ldots, \ell - 1\} \quad \text{arbitrary for} \quad \ell = m_j + 1, \ldots, M.$$

The proposed method gives a rigid solution of (AP) in the "most pessimistic case, i.e. if there is not a single pair $Q_{B_j}^j$, $Q_{B_i}^i$, $i \neq j$, of consistent distributions in the knowledge base.

In case of consistency, we could obviously obtain a solution with smaller M. But there are still good reasons for keeping this approach. First, we must have in mind the possibility of the pessimistic case. Further, with our approach we do not need to check the consistency which might be complicated and time-consuming. Finally, the pair-wise consistency does not in general yield the joint consistency. Thus, the graph approach in the following subsection could not be applied. (If we construct edges between pair of consistent distributions, then cliques would not represent jointly consistent sets of distributions.) On the other hand, there is always a possibility in frame of pre-processing to join some knowledge base distributions defined on overlapping sets (although this approach may not yield an improvement).

4.3. An algorithm for minimal covering

Let us construct a graph G on the set $\mathcal{M} = \{1, \ldots, M\}$ by the following principle

$$\langle k, \ell \rangle \in G \quad \text{iff} \quad B_k \cap B_\ell = \emptyset.$$

Then the solution of i) in the preceding subsection will be given by minimal covering of $\mathcal{M}$ by a system of cliques in the graph G. Since every clique can be extended to a maximal clique, and, on the other hand, maximal cliques can be restricted to their sub-cliques with preserving the covering requirement, we can deal only with the maximal cliques. Thus, let us denote by ψ the system of all maximal cliques in G. For finding ψ we can introduce a simple algorithm:

Step 1 Set $\psi = \emptyset$

Step 2 For $\langle k, \ell \rangle \in G$ do

2.1 Set $\mathsf{P} = \{k, \ell\}$

2.2 For $v \in \mathcal{M} \setminus \mathsf{P}$ do: if $\langle v, p\rangle \in G$ for every $p \in \mathsf{P}$ then $\mathsf{P} := \mathsf{P} \cup \{v\}$

Step 3 If $\mathsf{P} \notin \psi$ then $\psi := \psi \cup \{\mathsf{P}\}$.

Now, with $\psi = \{\mathsf{P}_1, \ldots, \mathsf{P}_m\}$ let $\phi \subset \psi$ be the minimal covering. We denote

$$z_i = I_{\mathsf{P}_i \in \phi}, \quad a_{i,k} = I_{k \in \mathsf{P}_i} \quad \text{for every } i = 1, \ldots, m \text{ and } k = 1, \ldots, M.$$

Then the problem can be rewritten in the following way

$$\min \sum_{i=1}^{m} z_i \quad \text{with} \quad \sum_{i=1}^{m} z_i\, a_{i,k} \gtrdot 0 \quad \text{for every } k = 1, \ldots, M.$$

But this is the "transportation problem" in linear programming, and it can be solved by linear programming methods (see e. g. [6], Chapter 9).

5. REPRESENTATION

5.1. Barycenter

Let us consider a finite collection of probability distributions

$$\boldsymbol{Q} = \{Q_S^1, \ldots, Q_S^M\}$$

on the space X_S. For some technical reasons we assume $\bigcap_{j=1}^{M} \operatorname{supp} Q_S^j \neq \emptyset$.

We would like to represent the collection by a single distribution. Naturally, the representing distribution should be found somewhere in the "centre of gravity". For this purpose the concept of barycenter seems appropriate. Following e. g. Perez [12], by the *barycenter* of the collection $\boldsymbol{Q}$ we understand a probability distribution $R_S^{\boldsymbol{Q}}$ satisfying

$$\max_{j=1,\ldots,M} H(R_S^{\boldsymbol{Q}} | Q_S^j) = \min_{R_S \in \mathcal{P}_S} \max_{j=1,\ldots,M} H(R_S | Q_S^j)$$

where $\mathcal{P}_S$ denotes the set of all probability distributions on X_S, and $H(\cdot|\cdot)$ is the relative entropy (I-divergence, Kullback–Leibler number) defined for a pair of discrete distributions P, Q by

$$H(P|Q) = \sum_{x \in \mathcal{X}} \log \frac{P(x)}{Q(x)}\, P(x) \geq 0 \quad \text{if } P \ll Q$$

and

$$H(P|Q) = +\infty \quad \text{otherwise.}$$

(Thus for $\bigcap_{j=1}^{M} \operatorname{supp} Q_S^j = \emptyset$ we would have $H(R_S|Q_S^j) = \infty$ for some $j = 1, \ldots, M$ and every $R_S \in \mathcal{P}_S$.)

In addition, let us denote by $\Gamma = \{\gamma = (\gamma_1, \ldots, \gamma_M) \in R^M; \sum_{j=1}^{M} \gamma_j = 1$ and $\gamma_i \geq 0$ for every $i = 1, \ldots, M\}$ the set of probability vectors on $\{1, \ldots, M\}$, and

$$Q_S^{\gamma} = \frac{\prod_{j=1}^{M} (Q_S^j)^{\gamma_j}}{c(\gamma)} \in \mathcal{P}_S$$

where $c(\gamma) = \sum_{x_S \in X_S} \prod_{j=1}^{M} (Q_S^j(x_S))^{\gamma_j}$ is the appropriate normalizing constant. (Here for $\bigcap_{j=1}^{M} \operatorname{supp} Q_S^j = \emptyset$ we would have $c(\gamma) = 0$.)

Lemma 1. For $\gamma \in \Gamma$ it holds

$$\sum_{j=1}^{M} \gamma_j \, H(Q_S^{\gamma}|Q_S^j) \leq \sum_{j=1}^{M} \gamma_j \, H(R_S|Q_S^j)$$

for every $R_S \in \mathcal{P}_S$.

Proof. We observe $\sum_{j=1}^{M} \gamma_j \, H(Q_S^{\gamma}|Q_S^j) = -\log c(\gamma)$, and, simultaneously,

$$0 \leq H(R_S|Q_S^{\gamma}) = \sum_{j=1}^{M} \gamma_j \, H(R_S|Q_S^j) + \log c(\gamma)$$

which proves the statement. □

Proposition 1. Let $c(\gamma^0) \leq c(\gamma)$ for every $\gamma \in \Gamma$. Then

$$Q_S^{\gamma_0} = R_S^{\boldsymbol{Q}}$$

is the barycenter.

Proof. We can follow Theorem 4.4.1 in Gallager [2] (see also Perez [12]) to obtain

$$H(Q_S^{\gamma^0}|Q_S^j) = -\log c(\gamma^0) \quad \text{whenever } \gamma_j^0 > 0,$$

and

$$H(Q_S^{\gamma^0}|Q_S^j) \leq -\log c(\gamma^0) \quad \text{for } \gamma_j^0 = 0.$$

Then, with the aid of Lemma 1, we finally have

$$\begin{aligned} \max_{j=1,\ldots,M} H(R_S|Q_S^j) &\geq \sum_{j=1}^{M} \gamma_j^0 \, H(R_S|Q_S^j) \geq \sum_{j=1}^{M} \gamma_j^0 \, H(Q_S^{\gamma^0}|Q_S^j) \\ &= -\log c(\gamma^0) = \max_{j=1,\ldots,M} H(Q_S^{\gamma^0}|Q_S^j) \end{aligned}$$

for arbitrary $R_S \in \mathcal{P}_S$. □

5.2. Modified practical solution

Thus, in order to find the barycenter of the collection $\boldsymbol{Q}$, it would be necessary to find $\gamma^0 \in \arg\min_{\gamma \in \Gamma} c(\gamma)$, which is again hardly possible. Therefore, it seems natural to choose $\gamma^0 \in \Gamma$ fixed and to minimize

$$\sum_{j=1}^{M} \gamma_j^0 \, H(R_S|Q_S^j)$$

instead of $\max_{j=1,\ldots,M} H(R_S|Q_S^j)$. Then, directly by Lemma 1, we have

$$Q_S^{\gamma^0} \in \operatorname*{arg\,min}_{R_S \in \mathcal{P}_S} \sum_{j=1}^{M} \gamma_j^0 \, H(R_S|Q_S^j)$$

as a solution.

Remark. Now, we may choose and interpret the coefficients $\{\gamma_j^0\}_{j=1,\ldots,M}$ as the weights describing "reliability" or "importance" assessed to the particular distributions Q_S^j, $j = 1, \ldots, M$. In this sense such substitute solution may even better correspond to the idea of "centre of gravity".

Let us recall that if $\gamma_j^0 > 0$ for every $j = 1, \ldots, M$ we obtain $Q_S^{\gamma^0}(x_S) = 0$ whenever $Q_{B_j}^j(x_{B_j}) = 0$ for some $j \in \{1, \ldots, M\}$. It means that if some configuration is forbidden by the knowledge base, it remains forbidden.

Moreover, let Q_S^j, $j = 1, \ldots, M$, be Gibbs distributions, i.e. $Q_S^j = P_S^{\Phi^j}$ for $j = 1, \ldots, M$. Then $Q_S^{\gamma^0}$ is also a Gibbs distribution, namely

$$Q_S^{\gamma^0} = P_S^{\sum_{j=1}^M \gamma_j^0 \Phi^j},$$

which also proves the convenience of such "geometric mean" approach in the present case. Obviously, $Q_S^{\gamma^0}$ is known up to the normalizing constant $c(\gamma^0)$ which is numerically unfeasible. Nevertheless, the simulation is possible as we shall see in the following section.

6. SIMULATION

6.1. Markov Chain Monte Carlo

For the sake of brevity we introduce the *Hamiltonian* $U_S^\Phi : X_S \to R$ given by

$$U_S^\Phi(x_S) = \sum_{A \in \mathcal{A}} \Phi_A(x_A)$$

for every $x_S \in X_S$ and some fixed potential Φ. Then the Gibbs distribution can be defined as

$$P_S^\Phi(x_S) = \frac{1}{Z_S^\Phi} \exp\{U_S^\Phi(x_S)\}.$$

For the sake of simplicity let us for now assume $U_S^\Phi > -\infty$, i.e. $P_S^\Phi > 0$. It is obvious that, still due to the number $|X_S|$ of all possible configurations, direct sampling is not feasible. We need approach that strictly avoids using terms and quantities involving the normalizing constants. Therefore, an iterative method based on constructing an appropriate Markov chain has been proposed. For sampling from the Gibbs distribution P_S^Φ we need a homogeneous Markov chain with transition probability matrix Q in order to satisfy $\nu Q^n \xrightarrow[n\to\infty]{} P_S^\Phi$ for any initial distribution ν. For the limit case $P_S^{\infty,\Phi} = \lim_{\beta\to\infty} P_S^{\beta\Phi}$, which is crucial for solving the optimization problems, we have to construct a non-homogeneous chain with $\nu Q_1 \ldots Q_n \xrightarrow[n\to\infty]{} P_S^{\infty,\Phi}$. In general, for this kind of methods the term *stochastic relaxation* is used, while the non-homogeneous case is known as *simulated annealing*.

6.2. Gibbs Sampler

For any $V \subset S$, let us set a *probability kernel* $\Pi_V : X_S \otimes X_S \to R$ by

$$\Pi_V(x_S; y_S) = P^{\Phi}_{V|S\setminus V}(y_V | x_{S\setminus V}) \cdot \delta(y_{S\setminus V} = x_{S\setminus V}).$$

We can easily verify that $P^{\Phi}_S \Pi_V = P^{\Phi}_S$, i.e. the Gibbs distribution P^{Φ}_S is invariant under the kernel Π_V. Let us emphasize that, under the action of Π_V, the configuration can be changed only inside the region V. Thus, in order to obtain an irreducible Markov chain, we need a composite kernel $Q = \Pi_{V_1} \cdots \Pi_{V_k}$, where $S = \bigcup_{i=1}^{K} V_i$ with small enough V_i, $i = 1, \cdots, k$. Usually, the elementary one-body sets are considered, namely

$$Q = \Pi_{s_1} \cdots \Pi_{s_{|S|}}$$

where $s_1, \ldots, s_{|S|}$ is some enumeration (*visiting scheme*) of the set S. Since now the Markov chain with the transition probability matrix Q is *ergodic*, we have $P^{\Phi}_S = \lim_{n\to\infty} \nu Q^n$ for any initial distribution ν.

Thus, we start from some initial configuration x^0_S. In the kth step, the configuration is updated at the site $s_{[k]}$, where $[k] = k \bmod |S|$, by sampling $\widetilde{x}_{s_{[k]}}$ from the conditional distribution $P^{\Phi}_{s_{[k]}|S\setminus\{s_{[k]}\}}\left(\cdot\,|x^{k-1}_{S\setminus\{s_{[k]}\}}\right)$, i.e. $x^{k-1}_S \mapsto \left(x^{k-1}_{S\setminus\{s_{[k]}\}}, \widetilde{x}_{s_{[k]}}\right) = x^k_S$. After $k = n|S|$ steps we have a sample from $\delta_{x^0_S} Q^n$. For large n we believe to have a sample from P^{Φ}_S.

Let us briefly note that some modifications, including a random visiting scheme, yield the same result. The algorithm was introduced by D. Geman and S. Geman [4] and was called the *Gibbs sampler*.

6.3. Metropolis Algorithm

Following an alternative idea introduced in Metropolis et al [10], we may directly set

$$Q(x_S; y_S) = R(x_S, y_S) \min\left(1, \frac{P^{\Phi}_S(y_S)}{P^{\Phi}_S(x_S)}\right) \quad \text{for } y_S \neq x_S$$

and $Q(x_S; x_S) = 1 - \sum_{z_S \neq x_S} Q(x_S; z_S)$, where $R(\cdot, \cdot)$ is a symmetric stochastic matrix. We can again verify that P^{Φ}_S is invariant under Q which is irreducible whenever R is irreducible. Then again $\lim_{n\to\infty} \nu Q^n = P^{\Phi}_S$ with any initial distribution ν.

The kth step of the *Metropolis algorithm* consists of two parts:

(I) A new configuration $\widetilde{x}^k_S$ is *proposed* by sampling from the distribution $R(x^{k-1}_S, \cdot)$.

(II) The proposed configuration is *accepted* for x^k_S at random with the probability equal to $\exp\left\{\min[0, U^{\Phi}_S(\widetilde{x}^k_s) - U^{\Phi}_S(x^k_S)]\right\}$.

A standard choice is $R(x_S, y_S) = \frac{1}{|S|(|X_s|-1)}$ if $x_s \neq y_s$ and $x_t = y_t$ for every $t \in S \setminus \{s\}$, and $R(x_S, y_S) = 0$ otherwise, i.e. we first choose uniformly at random a site $s \in S$ and then again uniformly a state $\widetilde{x}^k_s \neq x^{k-1}_s$. It is obvious that for large $|X_s|$ the Metropolis algorithm can be much faster to compare with the Gibbs sampler since it is not necessary to calculate all the local characteristics $P^{\Phi}_{s|S\setminus\{s\}}$.

Further generalization, namely the Metropolis–Hastings algorithm, is also possible. Here we set $Q(x_S; y_S) = R(x_S, y_S)\, M(x_S, y_S)$ for $y_S \neq x_S$, where the matrix M is chosen e.g. in order to preserve the detail balance equation $Q(y_S; x_S)\, P^{\Phi}_S(y_S) = Q(x_S; y_S)\, P^{\Phi}_S(x_S)$ for all x_S, $y_S \in X_S$, which again yields the invariance.

6.4. Simulated Annealing

Let us consider the probability distribution $P_S^{\infty,\Phi} = \lim_{\beta\to\infty} P_S^{\beta\Phi}$, i.e. $P_S^{\infty,\Phi}(x_S) = \frac{1}{|M^\Phi|} \cdot \delta(x_S \in M^\Phi)$ where

$$M^\Phi = \{x_S \in X_S; U_S^\Phi(x_S) = \max_{y_S \in X_S} U_S^\Phi(y_S)\}.$$

Since the parameter β has a standard physical interpretation as the inverse temperature, $P_S^{\infty,\Phi}$ is sometimes quoted as *zero temperature* distribution, or *ground state*.

Obviously, $\hat{x}_S$ is a sample from $P_S^{\infty,\Phi}$ if and only if $\hat{x}_S \in M^\Phi$. Therefore, the optimization problem $\max U_S^\Phi$, which is hardly solvable by any deterministic method, still can be solved by sampling from the distribution $P_S^{\infty,\Phi}$.

Since the support M^Φ of $P_S^{\infty,\Phi}$ is not known (otherwise there would be no problem) we cannot construct the respective kernel directly. Thus, we have to find a sequence $\{Q_n\}_{n=1}^\infty$ so that $P_S^{\infty,\Phi} = \lim_{n\to\infty} \nu Q_1 \cdots Q_n$. Due to the definition of $P_S^{\infty,\Phi}$, we shall define Q_n in order to satisfy $P_S^{\beta(n)\Phi} Q_n = P_S^{\beta(n)\Phi}$, namely

$$Q_n = \Pi_{s_1}^{\beta(n)} \cdots \Pi_{s_{|S|}}^{\beta(n)},$$

where, following the Gibbs sampler approach,

$$\Pi_{s_i}^{\beta(n)}(x_S; y_S) = P_{s_i|S\setminus\{s_i\}}^{\beta(n)\Phi}(y_{s_i}|x_{S\setminus\{s_i\}}) \cdot \delta(y_{S\setminus\{s_i\}} = x_{S\setminus\{s_i\}})$$

for every $i = 1, \ldots, |S|$ (see Section 6.2).

The inverse temperature $\beta(n)$ is assumed to be fixed during the nth sweep, and the sequence $\{\beta(n)\}_{n=1}^\infty$ with $\lim_{n\to\infty} \beta(n) = +\infty$ is called a *cooling schedule*. The choice of a proper cooling schedule is the crucial problem of the method. Let us introduce a standard theoretical result. It is known (cf. e.g. Theorem 5.2.1 in Winkler [13]) that for $\beta(n) \leq [|S|^{-1}\Delta^{-1}] \log n$ where

$$\Delta = \max_{s\in S} \{\sup |U_S^\Phi(x_S) - U_S^\Phi(y_S)|,\ x_{S\setminus\{s\}} = y_{S\setminus\{s\}}\}$$

is the maximal local fluctuation of U_S^Φ, we have $\lim_{n\to\infty} \mu Q_1, \ldots, Q_n = P_S^{\infty,\Phi}$ uniformly for all initial distributions μ.

Similar result can be also obtained for the Metropolis algorithm. The practical application of the *simulated annealing* method, as it is described above theoretically, is an art of its own. There are a number of problems connected to its implementation.

6.5. A generalization to non–positive and conditional distributions

Now, let us suppose $X_+(P_S^\Phi) \neq X_S$, i.e. P_S^Φ is not everywhere positive. A quick introspection of the above methods shows that with the same formal definitions in both Gibbs and Metropolis algorithms, the transition probability matrix Q acts on the set $X_+(P_S^\Phi)$ only. Namely, we can observe

$$Q(x_S; y_S) = 0 \quad \text{for } x_S \in X_+(P_S^\Phi) \text{ and } y_S \notin X_+(P_S^\Phi).$$

Therefore, with an initial distribution ν concentrated to $X_+(P_S^\Phi)$, we obtain a Markov chain with the state space $X_+(P_S^\Phi)$ and P_S^Φ as the stationary distribution, and all the above results remain valid.

Further, let us suppose $P_W^{\Phi}(\overline{x}_W) > 0$ for some $\overline{x}_W \in X_W$ and $\Phi = \{\Phi_A \in \mathcal{F}_A\}_{A \in \mathcal{A}}$. Then the conditional distribution

$$P^{\Phi}_{S \setminus W|W}(\cdot|\overline{x}_W)$$

is a Gibbs distribution on the state space $X_{S \setminus W}$ with the potential $\widetilde{\Phi}^{\overline{x}_W} = \{\widetilde{\Phi}_E^{\overline{x}_W} \in \mathcal{F}_E\}_{E \in \mathcal{E}}$ where $\mathcal{E} = \{A \cap (S \setminus W); A \in \mathcal{A}\}$ and

$$\widetilde{\Phi}_E^{\overline{x}_W}(x_E) = \sum_{A \in \mathcal{A}: A \cap (S \setminus W) = E} \Phi_A(x_E, \overline{x}_{W \cap A}).$$

Thus, all the above results on the simulation hold for the conditional distributions in the same way.

7. CONCLUDING REMARKS

7.1. Bootstrapping

Suppose we have a sequence of data

$$x_S^{(1)}, \ldots, x_S^{(N)},$$

and we are interested in an estimate $\widehat{P}_S$ of the underlying distribution P_S. And, unfortunately, the sample size N is so small that the empirical distribution would be completely unreliable.

Then we can calculate some small (e. g. only two-body) empirical marginals $\widehat{P}_{B_j}$, $j = 1, \ldots, K$, and understand them as the knowledge base in the above described method. Then we can produce a much larger sample with the aid of MCMC and calculate the empirical distribution $\widehat{\widehat{P}}_S$ from these re-sampled data. If the system $\{B_j\}_{j=1,\ldots,K}$ corresponds well to the true dependence structure and the approximation $Q_S^{\gamma^0}$ in Section 5.2 (with $\widetilde{Q}^j_{B_j} = \widehat{P}_{B_j}$ for $j = 1, \ldots, K$) fits well to the underlying distribution P_S, we can obtain a more reliable estimate. In this case we in fact follow the statistical principle of bootstrapping combined with a partial assumption on the probability model. On the other hand, e. g. the assumption on pair-wise interactions is not too much limiting.

7.2. Example

We applied the above described "bootstrapping" principle for checking the proposed method. We used a real data set $x_S^{(1)}, \ldots, x_S^{(N)}$ with $N \doteq 10^3$, $|S| = 35$, and $2 \leq |X_s| \leq 9$ for $s \in S$. Without any deeper study of the dependence structure the knowledge base was formally given by the system of all pair-wise empirical distributions

$$\widetilde{Q}_{\{s,t\}}; \quad s, t \in S, \ s \neq t,$$

i. e. $\mathcal{B} = \{B_j\}_{j=1}^{595} = \{\{s,t\}\}_{s,t \in S, s \neq t}$. Here we do not need to apply the algorithm in Subsection 4.3 to observe that there are approximately 34 partitions of S by the sets from $\mathcal{B}$ (in fact it is a bit more complicated due to the odd number $|S| = 35$).

For the sake of simplicity we can choose uniform γ^0, i.e. $\gamma_1^0 = \cdots = \gamma_{34}^0 = \frac{1}{34}$, to obtain

$$Q_S^{\gamma^0} = \left(\prod_{s,t \in S; s \neq t} \widetilde{Q}_{\{s,t\}} \right)^{\frac{1}{34}} \cdot \frac{1}{c(\gamma^0)}$$

as the representing distribution.

We used the Gibbs sampler for simulating a sequence

$$\widehat{x}_S^{(1)}, \ldots, \widehat{x}_S^{(n)}$$

with $n = 1\,500$ (for larger sample size the results remain unchanged).

Since the true underlying distribution P_S is unknown, we can compare the simulated results only with quantities obtained from the original data. And only a small-dimensional empirical marginals $\widetilde{Q}_B$ can be understood as good approximations of the true marginals P_B.

Therefore, a testing subset $B \subset S$ was chosen randomly with either $|B| = 3$ or $|B| = 4$, and the test statistics

$$H(\widehat{P}_B | \widetilde{Q}_B)$$

was calculated (see Subsection 5.1 for the definition of $H(\cdot|\cdot)$) where $\widetilde{Q}_B$ and $\widehat{P}_B$ are the empirical distributions calculated from the original and the simulated data, respectively.

The experiment was repeated 100 times for both $|B| = 3$ and $|B| = 4$. The results are contained in the following table. Let us remark that applying any statistical hypotheses testing method would not be useful because of the approximation error which cannot be excluded and which would cause rejecting the null hypothesis in many cases.

Table 1.

$H(\cdot\|\cdot)$	< 0.05	$0.05 - 0.1$	$0.1 - 0.15$	$0.15 - 0.2$	> 0.2
% for $\|B\| = 3$	65	16	5	9	5
% for $\|B\| = 4$	46	25	13	6	10

Thus the results are more-or-less for illustration but they do not seem to be completely unsatisfactory. It is obvious that by a more sophisticated choice of the knowledge base we could improve the results by decreasing the approximation error.

ACKNOWLEDGEMENT

The authors thank Radim Jiroušek for kindly providing the data.

(Received November 7, 1997.)

REFERENCES

[1] J. Besag: On the statistical analysis of dirty pictures (with discussion). J. Roy. Statist. Soc. Ser. B *48* (1986), 259–302.
[2] R. Gallager: Information Theory and Reliable Communication. J. Wiley, New York 1968.
[3] A. E. Gelfand and A. F. M. Smith: Sampling–based approaches to calculating marginal densities. J. Amer. Statist. Assoc. *85* (1990), 398–409.
[4] D. Geman and S. Geman: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. *6* (1984), 721–741.
[5] W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds.): Markov Chain Monte Carlo in Practice. Chapman and Hall, London 1996.
[6] G. Hadley: Linear Programming. Addison Wesley, Reading 1962.
[7] M. Janžura and S. Přeučil: An expert system based on the simulated annealing algorithm. In: WUPES 91, Prague 1991.
[8] S. L. Lauritzen: Graphical Models. University Press, Oxford 1996.
[9] F. Matúš: On iterations of average of I-projections. In: Highly Structured Stochastic Systems. Rebild, Denmark 1996.
[10] N. Metropolis, A. W. Rosenbluth M. N. Rosenbluth, A. H. Teller and E. Teller: Equations of state calculations by fast computing machines. J. Chem. Phys. *21* (1953), 1087–1092
[11] J. Moussouris: Gibbs and Markov random systems with constraints. J. Statist. Phys. *10* (1974), 1, 11–33.
[12] A. Perez: Barycenter of a set of probability measures and its application in statistical decision. In: Proceedings COMPSTAT 1984, Physica–Verlag, Wien 1984, pp. 154–159.
[13] G. Winkler: Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer–Verlag, Berlin 1995.

RNDr. Martin Janžura, CSc. and Mgr. Pavel Boček, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: janzura@utia.cas.cz, bocek@utia.cas.cz