Int. J. Appl. Math. Comput. Sci., 2006, Vol. 16, No. 2, 233–240

NODE ASSIGNMENT PROBLEM IN BAYESIAN NETWORKS

JOANNA POLANSKA*, DAMIAN BORYS*
ANDRZEJ POLANSKI*,**

* System Engineering Group, Silesian University of Technology
ul. Akademicka 16, 44–100 Gliwice, Poland
e-mail: {joanna.polanska,andrzej.polanski,damian.borys}@polsl.pl

** Department of Statistics, Rice University
PO Box 1892, MS-138, Houston, TX 77251, USA

This paper deals with the problem of searching for the best assignments of random variables to nodes in a Bayesian network (BN) with a given topology. Likelihood functions for the studied BNs are formulated, methods for their maximization are described and, finally, the results of a study concerning the reliability of revealing BNs' roles are reported. The results of BN node assignments can be applied to problems of the analysis of gene expression profiles.

Keywords: biostatistics, Bayesian networks, maximum likelihood, confidence intervals

1. Introduction

Bayesian networks (BNs) provide an economical and convenient representation of multidimensional joint probability distributions (Charniak, 1991; Jensen, 2001; Neapolitan, 2003; Pearl, 2000). The nodes of BNs correspond to random variables representing measurements or observations, and directed edges correspond to relations between these random variables. BNs can be constructed in two steps. The first step is to define a directed, acyclic graph (DAG) of causal relations between nodes. The second one consists in specifying conditional probability distributions corresponding to causal relations represented by the edges of the BN.

In the context of various applications of BNs, the following problems are often stated:

(i) Given observations, compute the probabilities of events, under the assumption of a known topology and known probability distributions in the BN.

(ii) Given observations and given a BN topology, estimate the parameters of conditional probability distributions.

(iii) Given observations, estimate the topology and probability distributions describing a BN.

Problems (i) and (ii) can be solved with the help of appropriate algorithms (Charniak, 1991; Heckerman, 1995; Jensen, 2001; Neapolitan, 2003; Pearl, 2000) and software (Murphy, 2005). Difficulties may arise for large numbers of nodes and edges. As for Problem (iii), there is a lot of interest and research in this direction (Chickering, 2002; Friedman, 1998; Liu and Desmarais, 1997; Pearl and Verma, 1991), motivated by the broad area of potential applications (Friedman, 2004; Friedman *et al.*, 2000; Ideker *et al.*, 2002; 2001; Jansen *et al.*, 2003; Segal *et al.*, 2001). However, the problem of estimating the topology of a BN is difficult, due to the large amount of uncertainty and serious computational complexity even for BNs of very moderate sizes.

In this paper we consider another problem, namely, that of assigning the observed random variables to the nodes of a Bayesian network model with a given topology. We call this problem the Bayesian network node assignment. It can be applied in the area of the analysis of DNA microarray data. DNA microarrays are often used to compare gene expression patterns of normal and cancer cells, cf. (Rhodes *et al.*, 2004), and it is believed that such comparisons will help in developing knowledge on the mechanisms of neoplasia. The characteristic property of the measurements of gene expressions patterns of cancer tissues in one experiment is, however, the coexistence of genes related to the processes of initiation of neoplasia (causes), as well as those related to the effects of neoplastic transformation. Both these classes, causes and effects, exhibit correlations with the status of the cell (normal/cancer). So they may be difficult to distinguish by simple statistics (Gadbury and Schreuder, 2003). In an effort towards constructing a test for distinguishing between causes and effects, we hypothesize the topology of a BN which models both the initiation of neoplastic transformation and its consequences. In the next step we cast

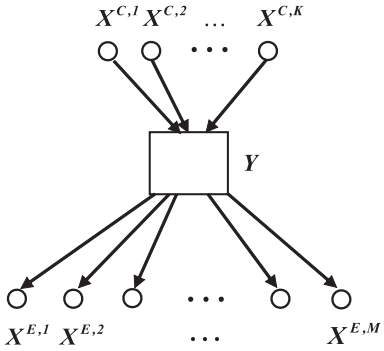

Fig. 1. Bayesian network model with cooperative causes. The causal relation $X^{C,1}, \ldots, X^{C,K} \rightarrow Y$ is modeled by a logistic function, and the influence of the state of Y on the values of $X^{E,1}, \ldots, X^{E,M}$ is modeled as changing values of expectations and variances of conditional distributions (see the text). In the case where the above BN is used to explain DNA microarray data, the continuous nodes represent base 2 logarithms of fluorescence signals at gene probes in a DNA microarray chip, and the binary node Y holds information on the status of the cells used in the experiment (normal "U" or cancer "T").

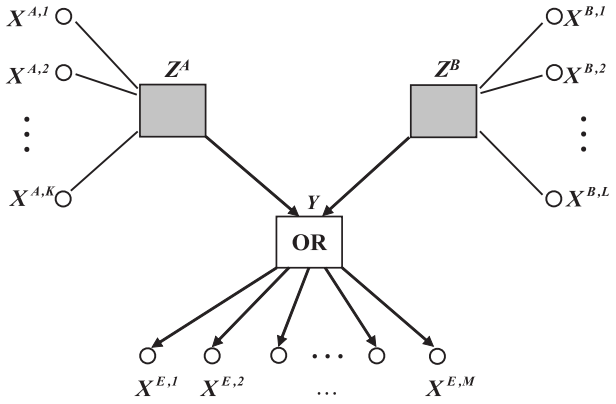

Fig. 2. Bayesian network model with two hypothetical (hidden) alternative causes, Z^A and Z^B. Again, the circles are continuous nodes and squares are binary nodes. When using the BN for DNA microarray data, the continuous nodes represent base 2 logarithms of fluorescence signals at gene probes in a DNA microarray chip, and the binary nodes represent biological processes. Y holds information on the status of the cells used in the experiment (normal "U" or cancer "T"). Z^A and Z^B are processes coded as "U" (inactive) and "T" (initiated), whose alternative leads to triggering the neoplastic transformation. Each of the processes Y, Z_1 and Z_2 has its specific effects shown in the diagram.

the problem of differentiating between the causes and effects of the neoplastic process as a BN node assignment. More results concerning the use of the BN assignment with DNA microarray data are presented in (Polanski *et al.*, 2005).

We recover the true assignment by the maximization of the likelihood over all possible assignments and over the parameters of a given BN. We estimate confidence levels for the obtained assignments using a Monte Carlo method which involves recording the numbers of errors in repeated reassignments. Studying the confidence of assigning variables to the nodes of a BN helps in verifying whether the assumed BN topology can satisfactory explain the observed data. We illustrate the method of estimating assignments by likelihood maximization with the use of two examples of BNs shown in Figs. 1 and 2. These examples show that the BN node assignment problem depends on the assumed topology of a BN. We also compare the maximum likelihood approach with other methods of estimating assignments.

2. BN Models

A Bayesian network B is a pair

$$B = (G, D), \tag{1}$$

where G is a directed acyclic graph (DAG) and D is a set of conditional probability distributions that corresponds to G. The nodes of the graphs of BNs correspond to random variables, signals, measurements or observations. The term "signal" or "random variable" is related to the node of the BN, and the names "measurements" or "observations" are used in the context of repetitive experiments.

In this paper we consider the topologies of BNs presented in Figs. 1 and 2. When using BNs as models for explaining DNA microarray measurements (Polanski *et al.*, 2005), we call the BN of Fig. 1 a scenario with cooperative causes, and the BN of Fig. 2 a scenario with two alternative causes. The nodes in BNs in Figs. 1 and 2 can be continuous, denoted by circles, or binary, marked by squares. Continuous nodes represent the levels of gene expressions and binary nodes represents biological processes, which can be triggered (active) or inactive. Shaded nodes denote random variables which are not observable and are called hidden or latent.

2.1. Cooperative Causes. In the BN model of cooperative causes in Fig. 1, the binary process Y is initiated by an additive combination of many events which equally contribute to the final level of the risk of triggering the process Y, and each single cause alone does not change the risk significantly. When using the BN of Fig. 1 to model microarray data (Polanski *et al.*, 2005), the binary node Y represents the state of the biological process of neoplastic transformation, which is triggered or inactive.

Within the classes, the nodes are exchangeable, and the assignment of random variables to the nodes of the Bayesian network is fully defined by deciding to which of the classes each random variable belongs. The probability

of initiating the binary process Y is set by a combined influence of the variables $X^{C,1},\ldots,X^{C,K}$ (causes). The random variables $X^{C,1},\ldots,X^{C,K}$ are assumed normal. The states of the binary variable Y are: "U" (inactive) and "T" (triggered). The conditional probability of initiating (triggering) the process Y is assumed to follow the logistic relation

$$P[Y=\text{“}T\text{”}] = P_L(x^{C,1},\ldots,x^{C,K}) = \frac{\exp[\beta(x^{C,1}+x^{C,2}+\cdots+x^{C,K})]}{1+\exp[\beta(x^{C,1}+x^{C,2}+\cdots+x^{C,K})]}, \quad (2)$$

where the index "L" in $P_L(\cdot)$ is for "logistic" and β is a parameter. The effects $X^{E,1},\ldots,X^{E,M}$ of the process Y are conditionally normal random variables depending on the state of Y, $y=$ "U" (inactive) or $y=$ "T" (triggered) in the following way:

$$\begin{aligned} X^{E,m} \mid y=\text{“}U\text{”} &\sim \mathcal{N}(\mu_U^m,\sigma_U^m),\\ X^{E,m} \mid y=\text{“}T\text{”} &\sim \mathcal{N}(\mu_T^m,\sigma_T^m),\\ m &= 1,2,\ldots,M. \end{aligned} \quad (3)$$

$\mathcal{N}(\mu,\sigma)$ denotes a normal distribution with the mean μ and standard deviation σ. Repeated measurements of the signals Y, $X^{E,1},\ldots,X^{E,M}$ are performed and the index "i" is used to distinguish between different measurements (realizations of random variables). The related notation for probability density functions is

$$p(x_i^{E,m} \mid y_i=\text{“}U\text{”}) = p_N(x_i^{E,m},\mu_U^m,\sigma_U^m) \quad (4)$$

and

$$p(x_i^{E,m} \mid y_i=\text{“}T\text{”}) = p_N(x_i^{E,m},\mu_T^m,\sigma_T^m),$$

where $m=1,2,\ldots,M$ are the labels of continuous signals, i denotes the index of a given observation, $p_N(x_i,\mu,\sigma)$ or $p_N(x_i)$ stand for the probability density function of a normal distribution. Lower case letters are used to denote realizations of random variables.

Based on the structure in Fig. 1 and Assumptions (1) and (2), one can write the following expression for the likelihood of the Bayesian network:

$$\begin{aligned} L = &\prod_{\substack{k\in C\\ i}} p_N(x_i^{C,k}) \prod_{i\in T} P_L(x_i^{C,1},\ldots,x_i^{C,K})\\ &\times \prod_{i\in U}\left[1-P_L(x_i^{C,1},\ldots,x_i^{C,K})\right]\\ &\times \prod_{\substack{m\in E\\ i\in U}} p_N(x_i^{E,m},\mu_U^m,\sigma_U^m)\\ &\times \prod_{\substack{m\in E\\ i\in T}} p_N(x_i^{E,m},\mu_T^m,\sigma_T^m). \end{aligned} \quad (5)$$

The classes of continuous nodes (variables) are "C" – causes and "E" – effects, and the states of the binary process Y are "U" – inactive and "T" – triggered.

2.2. Alternative Causes. In Fig. 2 we present a BN with alternative causes Z^A and Z^B, each having the ability to trigger the binary process Y. Again, in the case of microarray data, the binary node Y represents the state of the biological process of neoplastic transformation. Z^A and Z^B are (hypothetical) processes coded, as previously, as "U" – inactive and "T" – initiated, whose alternative leads to triggering the neoplastic transformation Y. Each of the processes Y, Z^A and Z^B has its specific effects. These are, respectively, $X^{E,1},\ldots,X^{E,M}$, $X^{A,1},\ldots,X^{A,K}$ and $X^{B,1},\ldots,X^{B,L}$. The binary nodes Z^A and Z^B are assumed hidden, and these processes are not observed.

Let us consider composite states defined by the triples Y, Z^A and Z^B. We assume that Z^A and Z^B cannot be active simultaneously (as alternative causes). Consequently, possible composite states of Y, Z^A and Z^B are U, U, U or T, T, U or T, U, T, respectively. The first one is observable since the state of Y is known and $Y=$ "U" implies $Z^A=U$ and $Z^B=U$. The remaining composite states T, T, U and T, U, T are hidden since by $Y=$ "T" one cannot know which of the processes Z^A and Z^B caused the triggering of Y.

The mechanism of changes in the expressions for $X^{E,1},\ldots,X^{E,M}$ secondarily to the ongoing neoplastic transformation is changing the conditional expectation and variance, described by the conditional distributions (3)–(4), as has already been explained. Analogous mechanisms are assumed for the altering expressions $X^{A,1},\ldots,X^{A,K}$ and $X^{B,1},\ldots,X^{B,L}$ secondarily to Z^A and Z^B:

$$\begin{aligned} p(x^{A,k} \mid z^A=U) &= p_N(x^{A,k},\mu_U^k,\sigma_U^k),\\ p(x^{A,k} \mid z^A=T) &= p_N(x^{A,k},\mu_T^k,\sigma_T^k), \end{aligned} \quad (6)$$

where $k=1,2,\ldots,K$, and

$$\begin{aligned} p(x^{B,l} \mid z^B=U) &= p_N(x^{B,l},\mu_U^l,\sigma_U^l),\\ p(x^{B,l} \mid z^B=T) &= p_N(x^{B,l},\mu_T^l,\sigma_T^l), \end{aligned} \quad (7)$$

where $l=1,2,\ldots,L$. The notation is analogous to that employed in (3) and (4).

Treating the composite states Y, Z^A and Z^B as a parameter which can assume three categories, U,U,U or T,T,U or T,U,T, we can write the following expression for the likelihood of the BN from Fig. 2:

$$L = L_{UUU}L_{TTU}L_{TUT}, \quad (8)$$

where

$$L_{UUU} = \prod_{\substack{m\in E\\ i\in U}} p_N(x_i^{E,m}, \mu_U^m, \sigma_U^m) \prod_{\substack{k\in A\\ i\in U}} p_N(x_i^{A,k}, \mu_U^k, \sigma_U^k) \times \prod_{\substack{l\in B\\ i\in U}} p_N(x_i^{B,l}, \mu_U^l, \sigma_U^l), \quad (9)$$

$$L_{TTU} = \prod_{\substack{m\in E\\ i\in T}} p_N(x_i^{E,m}, \mu_T^m, \sigma_T^m) \prod_{\substack{k\in A\\ i\in T}} p_N(x_i^{A,k}, \mu_T^k, \sigma_T^k) \times \prod_{\substack{l\in B\\ i\in U}} p_N(x_i^{B,l}, \mu_U^l, \sigma_U^l), \quad (10)$$

$$L_{TUT} = \prod_{\substack{m\in E\\ i\in T}} p_N(x_i^{E,m}, \mu_T^m, \sigma_T^m) \prod_{\substack{k\in A\\ i\in U}} p_N(x_i^{A,k}, \mu_U^k, \sigma_U^k) \times \prod_{\substack{l\in B\\ i\in T}} p_N(x_i^{B,l}, \mu_T^l, \sigma_T^l). \quad (11)$$

The classes of continuous nodes (variables) are "A", "B" and "E", and the states of the binary processes Y, Z^A and Z^B are "U" – inactive and "T" – triggered.

3. Maximizing Likelihood Functions

First we describe in detail the methods we used for maximizing the likelihood (5) of the BN from Fig. 1. Further we also describe the approach for the maximization of the likelihood (9) of the BN from Fig. 2. To avoid repetitions and for the sake of brevity, the description for the BN from Fig. 2 is less detailed.

Maximizing the likelihood function (5) of the BN from Fig. 1 over the assignments of continuous signals to the nodes of the BN is a combinatorial optimization problem in which, theoretically, the structure with the maximal likelihood can be found by going through all assignments. However, going through all assignments can be prohibitive due to a large number of possible variants. For example, if 50 continuous random variables are to be assigned to nodes of two types, "cause" or "effect", then there are 2^{50} possible assignments. For that reason, when maximizing the likelihoods of BNs we used the Metropolis-Hastings algorithm (Gilks *et al.*, 1996; Metropolis *et al.*, 1953), where different assignments were sampled with frequencies depending on their likelihoods.

The use of the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953) for maximizing the likelihood function (4) with switching between assignments involves assuming an initial assignment of the BN, setting "current assignment" ← "initial assignment", and then repeating the following steps:

1. Obtain a "modified assignment" by introducing a random change in the "current assignment".
2. Make the substitution, "current assignment" ← "modified assignment", with the probability

$$P = \min\left(1, \frac{L(\text{"modified assignment"})}{L(\text{"current assignment"})}\right). \quad (12)$$

By L("assignment") we mean the likelihood function computed as in (5), which relies on a hypothesis concerning assigning a random variable to the nodes in the BN.

3.1. Coding and Random Modification of Assignments. In computer memory, the structure of a BN can be stored as an N-dimensional vector STRUC with entries coding the types of nodes. N is the number of random variables and the codes can be as follows: Nodes are coded as 1 (cause) and 2 (effect). For example, STRUC $=$ $[2\ 2\ 2\ 1\ 1\ 1\ 1]$ means that the BN in Fig. 1 has 7 continuous nodes and the random variables numbered from 1 to 3 are assigned to three effects of the process Y, while the random variables numbered from 4 to 7 are assigned to four causes. In other words, the node classes are causes $C = [4\ 5\ 6\ 7]$ and effects $E = [1\ 2\ 3]$. Since STRUC defines an assignment, for the likelihood in (4) we also use the notation

$$L = L(\text{STRUC}). \quad (13)$$

Using the notation (6) in the Metropolis-Hastings algorithm, in (5), the assignments can be defined by their vectors STRUC,

$$L(\text{"current assignment"}) = L(\text{current STRUC}),$$

$$L(\text{"modified assignment"}) = L(\text{modified STRUC}).$$

Since we assume a fixed topology of the BN, we keep fixed proportions between the types of nodes, causes and effects. Changing assignments in the algorithm of likelihood maximization involves a random choice of two indices of STRUC elements and then swapping the entries of STRUC corresponding to these indices.

3.2. Parameter Fitting. After an assignment (the STRUC vector) has been defined, the probabilities in (5) are computed using (4). However, the values of the likelihoods of BNs depend on parameter values, means and variances of normal and conditionally normal distributions, as well as on one parameter, β, of the logistic curve (2).

Optimizing the location and scale parameters of (conditionally) normal distributions is done by the standard expressions

$$\hat{\mu} = \frac{1}{I}\sum_{i=1}^{I} x_i, \quad \hat{\sigma} = \frac{1}{I-1}\sum_{i=1}^{I}(x_i - \hat{\mu})^2, \quad (14)$$

where x_i, $i = 1, 2, \ldots, I$ stand for repeated measurements. Fitting the parameter β in the logistic relation can, in principle, be done with the use of the maximum likelihood method by Newton-Raphson iterations. However, for the repetitions of observations which we used in our computations (30–100), the estimators of β have rather large variations and, therefore, we work with models with $\beta = 1$.

3.3. Maximizing the Likelihood of the BN with Alternative Causes. If for all observations we know the category U, U, U or T, T, U or T, U, T to which they belong, the likelihood (9) can readily be optimized by deciding, for all nodes, to which of the classes A, B or E they should be assigned. The parameters of conditionally normal distributions are again estimated by the standard formulas (14), and the decision is made based on maximizing the likelihood over three possible assignments (A, B or E). Maximizing over the hidden categories T, T, U or T, U, T becomes again a combinatorial optimization problem which can be solved by Metropolis-Hastings recursions. Let us note that the maximization of the likelihood method allows estimating K, L and M, i.e., the numbers of nodes in the classes A, B and E.

3.4. Identifiability. Reconstructing the roles of the nodes of BNs raises the question of identifiability, which is related to the theory of observational equivalence of BNs. Two DAGs, G and G', are observationally equivalent if for every Bayesian network $B = (G, D)$ there exists a Bayesian network $B' = (G', D')$ such that B and B' define the same probability distribution, and vice versa. The well-known theorem (e.g., (Pearl and Verma, 1991), p. 19, Th. 1.2.8) states that two DAGs are observationally equivalent iff they have the same skeletons and the same v-structures. Graphical explanations of the notions of the skeleton and v-structure are shown in Fig. 3. Observational equivalence concerns general, non-parametric conditional probability distributions describing the edges of BNs. When probability distributions are restricted to parametric classes, theorems on observational equivalence may not be true. However, identifying BN structures by detecting parametric classes of the edges is rather an artificial idea. In contrast, observational equivalence is a basic property of BN structures and has rather important consequences in inference based on models of BNs (Chickering, 2002).

Using the concept of observational equivalence, we call a DAG G identifiable if every DAG, G', obtained from G by reversing arrows is not observationally equivalent to G. In other words, a DAG is identifiable if its arrows cannot be reversed without changing the probability distributions of random variables corresponding to the nodes of the related BN. In the case where the number of causes is greater than or equal to 2, the DAG in Fig. 1 is identifiable. Since we understand the cooperative mechanism as many causes contributing to the final risk, the condition that their number is greater than 2 is always satisfied.

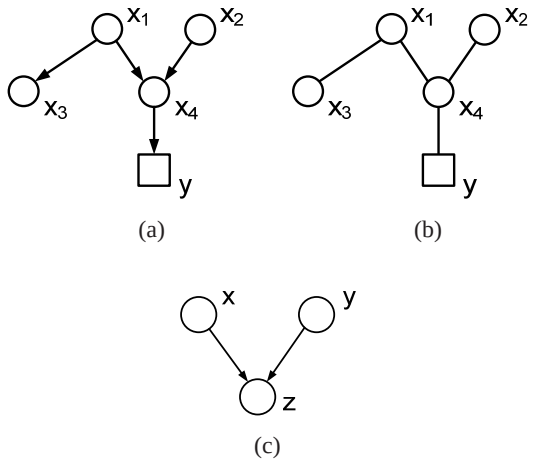

Fig. 3. Definitions related to the concept of observational equivalence of DAGs: (a) an exemplary DAG, (b) its skeleton obtained by the ignoring directions of the arrows in the DAG, (c) the v-structure, $X \rightarrow Z \leftarrow Y$, being a fragment of the DAG. In the definition of the v-structure it is required that there are no direct edges between X and Y.

The Bayesian network from Fig. 2 is not identifiable in the sense specified above. Based on the relation of observational equivalence, the scenario which we study can be represented by a partially directed graph (PDAG), (David and Nagaraja, 2003; Pearl and Verma, 1991; Pe'er *et al.*, 2001), where only some of the causal dependencies can be inferred. This is seen in Fig. 2, where directed edges represent causal relations which can be inferred from data, and undirected edges represent relations whose directions can be reversed within equivalence classes. In the BN in Fig. 2 the processes Z^A and Z^B are causes of Y. The causal dependence between $X^{A,1}, \ldots, X^{A,K}$, $X^{B,1}, \ldots, X^{B,L}$ and Y is not resolved.

4. Confidence Intervals for Assignments

The assignment with the highest likelihood is chosen as closest to the truth. However, for some BNs the obtained assignment may be very reliable while for others it may be distorted by large errors. Therefore, an important issue is estimating the probabilities of true or false assignments, as we show in this section. We discuss the case where we know the true assignment and we want to estimate the probabilities of its correct or erroneous recovery by an algorithm of likelihood maximization. We measure how close a given assignment is to the true assignment by counting misassignments defined by how many nodes are swapped between their true classes. The most straight-

Table 1. Results of 100 runs of 2500 steps of the Metropolis-Hastings algorithm for the maximization of the likelihoods of the Bayesian networks BNC1 and BNC2 described in the text.

Number of misassignments	0	1	2	3	4
Times encountered (BNC1)	6	43	32	15	4
Times encountered (BNC2)	90	10	0	0	0

Table 2. Results of 100 runs of 2500 steps of the Metropolis-Hastings algorithm for the maximization of the likelihoods of the Bayesian networks BNC1 and BNC2 described in the text.

Number of misassignments	0	1	2	3	4	5	6	7
Times encountered (BNA1)	0	0	0	20	13	27	27	13
Times encountered (BNA2)	92	8	0	0	0	0	0	0

forward approach to estimating the probabilities of true versus false assignments is by repeated simulations. We generate randomly observations according to the assumed BN model and, using the generated observations, we estimate the assignment, as described previously. In the case where we only have an estimate of an assignment at our disposal, a reasonable choice is to treat the estimate as the true assignment and to use Monte Carlo reassignments under this hypothesis.

Approaching the problem of predicting the probabilities of misassignments analytically is rather not feasible. It is possible to derive approximate methods. We developed a method of the analysis of assignment errors based on additive approximations which will be presented elsewhere.

Using the method of repeated simulations, we analyzed two examples of BNs, from Fig. 1. We denote them by BNC1 and BNC2. Both BNC1 and BNC2 contain 10 causes $X^{C,1},\ldots,X^{C,10}$ and 20 effects $X^{E,1},\ldots,X^{E,20}$. In both BNC1 and BNC2, the random variables $X^{C,1},\ldots,X^{C,10}$ are normally distributed with $\mu=0$ and $\sigma=1$, and the random variables $X^{E,1},\ldots,X^{E,20}$ are conditionally normally distributed, as described in (2), with $\sigma_U^m=\sigma_T^m=1$. BNC1 and BNC2 differ in the values μ_U and μ_T. For BNC1 we assumed $\mu_U^m=-0.8$, $\mu_T^m=0.8$, $m=1,2,\ldots,20$, and for BNC2 we have taken $\mu_U^m=-1.5$, $\mu_T^m=1.5$, $m=1,2,\ldots,20$. This assumption means that triggering the process Y changes the expectation of the conditional distribution of $X^{E,m}$ from $\mu_U^m=-0.8$ to $\mu_T^m=0.8$ in BNC1 and from $\mu_U^m=-1.5$ to $\mu_T^m=1.5$ in BNC2. The number of repeated measurements, in both BNs, was taken as 30. We repeated the likelihood optimization procedure 100 times of 2500 steps of the Metropolis-Hastings algorithm described in Section 3, and the obtained the results are presented in Table 1.

We also analyzed, by the method of repeated simulations, two instances of BNs, from Fig. 2. We denote them by BNA1 and BNA2. For both BNA1 and BNA2, we assumed that the classes A, B and E contain $K=10$, $L=10$ and $M=20$ nodes, respectively. In both BNA1 and BNA2 the true distributions of $X^{A,1},\ldots,X^{A,K}$, $X^{B,1},\ldots,X^{B,L}$ and $X^{E,1},\ldots,X^{E,20}$ are conditionally normally distributed, as described in the expressions (4), (6) and (7). For BNA1 we assumed $\mu_U^m=-0.8$, $\mu_T^m=0.8$, $k=1,2,\ldots,10$, $l=1,2,\ldots,10$, $m=1,2,\ldots,20$ in (4), (6) and (7), and for BNA2 we set $\mu_U^m=-1.5$, $\mu_T^m=1.5$, $k=1,2,\ldots,10$, $l=1,2,\ldots,10$, $m=1,2,\ldots,20$. The number of repeated observations in both BNs was taken as 30. For each observation we chose randomly the categories U,U,U or T,T,U or T,U,T with the probabilities $P(U,U,U)=0.5, P(T,T,U)=0.25$ and $P(T,U,T)=0.25$. We repeated the likelihood optimization procedure 100 times of 2500 steps of the Metropolis-Hastings algorithm, over the assignments of observations to the hidden categories T,T,U and T,U,T, and we present the obtained results in Table 2.

5. Remarks on Methods of Estimating Node Assignments

From the results of Tables 1 and 2, one can see that recovering assignments for BNC2 and BNA2 can be done more reliably than for BNC1 and BNA1. This results from the fact that differences between the distributions of the classes of nodes, causes and effects are larger in BNC2 and BNA2 and smaller in BNC1 and BNA1. So, intuitively, confusing the classes of nodes is more probable in BNC1 and BNA1 and less probable in BNC2 and BNA2.

The entries in Tables 1 and 2 corresponding to BNC1 and BNA1 reveal the remarkable fact that recovering a true assignment can be less probable than recovering an erroneous one. This fact follows from the effect of the maximization of the likelihood over many assignments. Define q as the number of misassignments. For example, $q=0$ means a true assignment and $q=1$ represents one cause switched with one effect. When comparing $q=0$ with $q=1$ in BNC1, the true assignment $q=0$ is "competing" not with one assignment $q=1$ but rather with all $K\ M$ possible different assignments, all leading to $q=1$, resulting from swapping $X^{C,1}$ with $X^{E,1}$, $X^{C,1}$ with $X^{E,2}$ ... and so forth. The likelihood of the true assignment is compared with the maximum over all of above combinations of distributions (David and Nagaraja, 2003). The maximization of the likelihood over even more structures takes place when comparing the true assignment $q=0$ to assignments leading to $q=2$, etc. An analogous mechanism acts in BNA1.

The above mechanism of a systematic error resulting from maximizing the likelihood over assignments suggests that care must be taken when approaching various problems of node assignment and indicates a need for comparing the maximum likelihood method with other methods of estimating assignments. A method often used in estimating the topologies of BNs is computing and comparing correlations between nodes. For the BN with cooperative causes from Fig. 1, the assignment of the random variable X to the classes "cause" and "effect" could be done by computing the correlation coefficient between X and Y and assigning X as a "cause" when the correlation coefficient is low and as an "effect" when it is high. Following the suggestion of an anonymous reviewer of the first version of the manuscript, we have compared the correlation approach to the maximum likelihood method. The conclusion is that, for the BN with cooperative causes from Fig. 1, the assignment by the values of the correlation coefficient is generally more efficient than that by maximizing the likelihood over assignments, in the sense that, on the average, it leads to assignments closer to the true one. However, for the BN with alternative causes from Fig. 2, the situation is opposite. By using correlations between X and Y, it is rather hopeless to stratify nodes into three classes A, B and E. Even when differentiating between two classes $A \cup B$ and E, correlations are, on the average, less efficient than the maximum likelihood. These observations suggest that more research is needed to establish more reliable and robust methods for the node assignment in BNs.

6. Conclusions

Estimating the topology of a BN based on the observation of realizations of random variables is a difficult problem, and the results reported in the literature (Chickering, 2002; Friedman, 1998; Liu and Desmarais, 1997; Pearl and Verma, 1991) concern mostly BNs of rather small sizes. Also very little is known about the probability distributions of errors in estimating BN topologies.

We presented some results related to the problem where the topology of a BN is known and uncertainty is in an unknown assignment of the observed random variables to the BN nodes. We based our approach on estimating the assignment via likelihood maximization. Simulation results show that node assignments can be quite reliably recovered based on a very moderate number of repeated observations (about 30). Estimation quality increases with increasing differences between the distributions of the classes of nodes. Estimating the probabilities of errors in node assignments allows evaluating the reliability of conclusions drawn based on the analyzed data.

Our research leads to understanding the probability distribution of the error committed when estimating the assignments of the observed random variables to the nodes of BNs by using the maximum likelihood approach. The results of simulations illustrate the mechanisms leading to misassignments in problems of the reconstruction of BNs, related to maximizing the likelihood over many assignments.

The development of biological knowledge, e.g., of mechanisms leading to neoplasia can involve constructing hypothetical cause-effect dependencies where the scenario is known but the roles of actors (genes) remain to be identified. Such problems can be formalized with the use of the concept of Bayesian networks and examined by using the methods presented in this paper.

Acknowledgments

The authors are grateful to the anonymous reviewer for his helpful comments. This paper was supported by the Ministry of Science and Higher Education, grant no. 3 T11F 010 29.

References

Charniak E. (1991): *Bayesian networks without tears*. — AI Magazine, Vol. 12, No. 4, pp. 50–63.

Chickering D.M. (2002): *Learning equivalence classes of Bayesian-network structures*. — J. Mach. Learn. Res., Vol. 2, No. 3, pp. 445–498.

David H.A. and Nagaraja H.N. (2003): *Order Statistics*. — Hoboken, New Jersey: Wiley.

Friedman N. (1998): *The Bayesian structural EM algorithm*. — Proc. 14-th Conf. *Uncertainty in Artificial Intelligence*, Madisin, Wisconsin, USA, pp. 129–138.

Friedman N. (2004): *Inferring cellular networks using probabilistic graphical models*. — Science, Vol. 303, No. 5659, pp. 799–805.

Friedman N., Linial M., Nachman I. and Pe'er D. (2000): *Using Bayesian networks to analyze expression data*. — J. Comput. Biol., Vol. 7, Nos. 3–4, pp. 601–620.

Gadbury G.L. and Schreuder H.T. (2003): *Cause-effect relationships in analytical surveys: An illustration of statistical issues*. — Env. Monit. Assess., Vol. 83, No. 3, pp. 205–227.

Gilks W.R., Richardson S. and Spiegelhalter D.J. (1996): *Markov Chain Monte Carlo in Practice*. — London: Chapman and Hall.

Heckerman D. (1995): *A tutorial on learning with Bayesian networks*. — Tech. Rep., MSR-TR-95-06, available at: `ftp://ftp.research.microsoft.com/pub/tr/tr-95-06.pdf`

Ideker T., Thorsson V., Ranish J.A., Christmas R., Buhler J., Eng J.K., Bumgarner R., Goodlett D.R., Aebersold D.R. and Hood L. (2001): *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network*. — Science, Vol. 292, No. 5518, pp. 929–934.

Ideker T., Ozier O., Schwikowski B. and Siegel A.F. (2002): *Discovering regulatory and signaling circuits in molecular interaction networks*. — Bioinf. Vol. 18, Suppl. 1, No. 90001, pp. S233–S240.

Jansen R., Yu H., Greenbaum H., Kluger Y., Krogan N.J., Chung S., Emili S., Snyder M., Greenblatt J.F. and Gerstein M. (2003): *A Bayesian networks approach for predicting protein – Protein interactions from genomic data*. — Science, Vol. 302, No. 5644, pp. 449–453.

Jensen F.V. (2001): *Bayesian Networks and Decision Graphs*. — New York: Springer.

Murphy K. (2005): *Bayes net toolbox for matlab*. — Available at: `http://bnt.sourceforge.net/`

Liu J. and Desmarais M.C. (1997): *A method of learning implication networks from empirical data: Algorithm and Monte-Carlo simulation-based validation*. — IEEE Trans. Knowl. Data Eng., Vol. 9, No. 6, pp. 990–1004.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953): *Equations of state calculations by fast computing machines*. — J. Chem. Phys., Vol. 21, No. 6, pp. 1087–1092.

Neapolitan R.E. (2003): *Learning Bayesian Networks*. — Upper Saddle River, NJ: Prentice Hall.

Pearl J. (2000): *Causality: Models, Reasoning, and Inference*. — Cambridge, MA: Cambridge University Press.

Pearl J. and Verma T.S. (1991): *A theory of inferred causation*, In: Principles of Knowledge Representation and Reasoning, (J.A. Allen, R. Fikes and E. Sandewall, Eds.). — San Mateo: Morgan Kaufmann.

Pe'er D., Regev A., Elidan G. and Friedman N. (2001): *Inferring subnetworks from perturbed expression profiles*. — Bioinf., Vol. 17, Suppl. 1, No. 90001, pp. S215–S224.

Polanski A., Polanska J., Jarzab M., Wiench M. and Jarzab B., (2005): *Inferring cause – effect relations from gene expression profiles of cancer versus normal cells*. — Tech. Rep., available at: `http://web.zis.ia.polsl.gliwice.pl/publikacje/projekty/technical_report.pdf`

Rhodes D.R., Yu J., Shanker K., Deshpande N., Varambally R., Ghosh R., Barrette T., Pandey A. and Chinnaiyan A.M. (2004): *ONCOMINE, A cancer microarray database and integrated data mining platform*. — Neoplasia, Vol. 6, No. 1, pp. 1–6.

Segal E., Taskar B., Gasch A., Friedman N. and Koller D. (2001): *Rich probabilistic models for gene expression*. — Bioinf., Vol. 1, No. 1, pp. 1–10.

Received: 20 September 2005
Revised: 14 March 2006