Model similarity and rank-order based classification under the assumption of positive association

by

Sung-Ho Kim and Geon Youp Noh

BK21 Research Report

09 - 07

May 19, 2009

DEPARTMENT OF MATHEMATICAL SCIENCES

The second stage of BK21 Fostering A World Class Talent

KAIST 한국과학기술원 Korea Advanced Institute of Science and Technology

Model Similarity and Rank-Order Based Classification under the Assumption of Positive Association

Sung-Ho Kim and Geon Youp Noh[1]
KAIST

Abstract: Suppose that we rank-order the conditional probabilities for a group of subjects that are provided from a Bayesian network model of binary variables. The conditional probability is the probability that a subject has a certain attribute given an outcome of some other variables and the classification is based on the rank-order. Under the condition that the class sizes are equal across the classes and that all the variables in the model are positively associated with each other, we compare the classification results between models which share the same model structure. Simulation results indicate that the agreement level of the classification between models is considerably high with the exact agreement for about half of the subjects or more and the agreement up to level 1 for about 90% or more.

Keywords: Agreement level; Bayesian network; Conditional probability; Model similarity; Positive association.

1 Introduction and Problem

Consider a problem of rank-ordering a group of subjects according to the conditional probability that a subject possesses a certain attribute given the outcome of a set of covariates. If we need to classify them into one of L classes according to the rank-order, it becomes a classification problem. This is an example of discretizing continuous values by rank-ordering, which takes place in a variety of performance assessment in education, sports, finance, etc. The rank-orders may be discrepant between raters due to different standards of thresholds among others. We will consider a rank-order based classification problem using the probability values which are produced from a Bayesian network (BN) model(Pearl 1988; Jensen 1996)

[1]Authors' Addresses: Sung-Ho Kim, Department of Mathematical Sciences, KAIST, Daejeon, 305-701, S. Korea, e-mail: sung-ho.kim@kaist.edu; Geon Youp Noh, Department of Mathematical Sciences, KAIST, Daejeon, 305-701, S. Korea, e-mail: dryleaf0@kaist.ac.kr.

Literature abounds on agreement measures between raters or classifiers (Cohen 1960; Spitzer, Cohen, Fleiss, and Endicott, 1967; Fleiss 1981; Messatfa 1992; Pires and Branco 1997; Albatineh, Niewiadomska-Bugaj, and Mihalko 2006; Brusco and Steinly 2008), where the agreement measure accounts for the level of agreement relative to the level of agreement by chance. The value of the measure is between 0 and 1, 1 indicating a perfect agreement between raters. This kind of agreement measures are very useful when we compare raters in a relative sense. In this paper, we are interested in a detailed description of agreement such as proportions of perfect agreement, agreement up to level 1, etc.

BNs are useful in representing graphically the relationships among the variables which are interpretable in generic terms as causal. In this paper, we will consider BN models of binary variables each taking on 0 or 1. When all the variables in the model are categorical with finite number of category levels, we can compute conditional probabilities of a set of variables given another set of variables in the model by applying the method called evidence propagation method (Lauritzen and Spiegelhalter 1988) and computer programs such as HUGIN (Andersen, Jensen, Olesen, and Jensen 1989), ERGO (Noetic Systems 1991), and MSBNx (MSBNx 2001) are available for the computing.

Suppose that we rank-order the conditional probabilities for a group of subjects that are provided from a BN model. The conditional probability is given in the form of $P(U = 1|\mathbf{X} = \mathbf{x})$ where $\mathbf{X}$ is a random vector. If we classify the subjects in such a way that the class sizes are equal across the classes, the classification is analogous to assigning rank-order-based grade scores to students. If two BN models provide the conditional probabilities for the same group of subjects and the rank-orders of the probability values are similar to each other between the two models, then the agreement level between the two classification results will be very high. We will investigate the agreement level between the BN models which satisfy some conditions as described in sections 4 and 5.

We will introduce a notion of similarity between BN models whose model structures are the same. The only difference between them is in the marginal probability or the conditional probability of a variable, X say, given the variables at its parent nodes, i.e., the nodes at the arrow tails of X, in the BN (see Figure 1). Assuming that the variables in a model are positively associated (Holland and Rosenbaum 1986; Junker and Ellis 1997) with each other, we will investigate agreement levels in classification between a pair of BN models, where one model is a similar model of the other.

This investigation is of practical value since it will provide us with a BN model whose overall agreement level with a group of BN models is relatively high under the assumption of positive association among the variables that are involved in the model. As for a BN model, the positive association between variables can be interpreted as a positively causal relationship. In cognitive science, we assume that knowledge states and task performances are positively causally related (Mislevy 1994) and we can find many such examples in biological, medical, and behavioral sciences among others.

This paper is organized in 6 sections. Section 2 gives a brief review on positive association among binary variables. Section 3 then introduces the notion of similarity between BN models and proposes a BN model which is most similar to a given BN model under some condition. In section 4, we describe a method of computing agreement levels of classifications between a pair of BN models and then results of a simulation experiment are presented. The simulation experiment extends in section 5 to the case where the comparison is made between a set of BN models and a BN model which satisfy a certain condition. Section 6 concludes the paper with some summarizing remarks.

2 A Brief Review on Positive Association

We will use U for unobservable variables and X for observable variables. For a pair of n-vectors $\mathbf{u}$ and $\mathbf{v}$ of the same length, we write $\mathbf{u} \preceq \mathbf{v}$ when $u_i \leq v_i$ for $i = 1, \cdots, n$, and write $\mathbf{u} \prec \mathbf{v}$ if $\mathbf{u} \preceq \mathbf{v}$ and $u_i < v_i$ for at least one $i = 1, \cdots, n$.

In a BN such as the graph in Figure 1, if a pair of nodes a and b are connected by an arrow with the arrow heading towards b from a, we call node a a parent node of node b and call node b a child node of node a. If a node does not have any parent node, it is called a root node. U_1 is the only root node in the figure.

Theorem 1. *Let U and X be binary variables, taking on the values 0 or 1. If*

$$0 < P(U = 1) < 1 \quad \text{and} \quad 0 < P(X = 1) < 1 \tag{1}$$

then the following two inequalities are equivalent:

$$P(X = 1|U = 0) < P(X = 1|U = 1) \tag{2}$$

$$P(U = 1|X = 0) < P(U = 1|X = 1). \tag{3}$$

Proof. See Theorem 1 in Kim (2005). □

Expression (2) can be re-expressed as

$$\frac{P(X = 1|U = 1)P(X = 0|U = 0)}{P(X = 0|U = 1)P(X = 1|U = 0)} > 1,$$

which means that U and X are positively associated.

Theorem 1 can be extended to random vectors as the theorem below.

Theorem 2. *Let* $\boldsymbol{X} = (X_1, \cdots, X_I)$ *and* $\boldsymbol{U} = (U_1, \cdots, U_K)$ *where all the* X_i*'s and* U_k*'s are binary, taking on 0 or 1. Then the following two statements are equivalent.*

(i) *For* $i = 1, 2, \cdots, I$,

$$P(X_i = 1|\mathbf{u}) < P(X_i = 1|\mathbf{v}), \quad \text{when} \quad \mathbf{u} \prec \mathbf{v}. \tag{4}$$

(ii) *For* $k = 1, 2, \cdots, K$,

$$P(U_k = 1|\mathbf{x}) < P(U_k = 1|\mathbf{y}), \quad \text{when} \quad \mathbf{x} \prec \mathbf{y}. \tag{5}$$

The strict inequality(<) in both Eqs.(4) and (5) may be replaced by the plain inequality (≤).

Proof. See Theorem 2 in Kim (2005). □

Inequality (4) is equivalent to that

$$\frac{P(\mathbf{v}|X_i = 1)}{P(\mathbf{u}|X_i = 1)} > \frac{P(\mathbf{v})}{P(\mathbf{u})} > \frac{P(\mathbf{v}|X_i = 0)}{P(\mathbf{u}|X_i = 0)}.$$

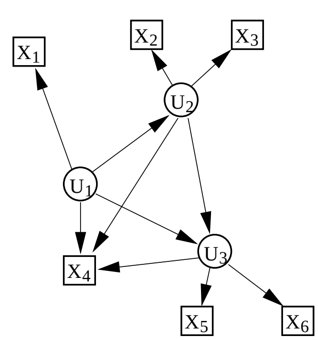


Figure 1: A BN where U variables are latent and X variables observable.

This inequality says that $\mathbf{U} = \mathbf{v}$ is more likely than $\mathbf{U} = \mathbf{u}$ when $X_i = 1$ than when $X_i = 0$. Furthermore, we can compare the likeliness of $U_k = 1$ for individual U_k's as in Theorem 2.

When the $<$ and $\prec$ in expression (4) are replaced by $\leq$ and $\preceq$, respectively, we have

$$P(X_i = 1|\mathbf{u}) \leq P(X_i = 1|\mathbf{v}), \quad \text{when} \;\; \mathbf{u} \preceq \mathbf{v}. \tag{6}$$

This expression actually means that the conditional distribution of X_i given $\mathbf{U} = \mathbf{v}$ is stochastically larger than that of X_i given $\mathbf{U} = \mathbf{u}$. Holland and Rosenbaum (1986) discuss properties concerning positive association among the X variables when the X variables are conditionally stochastically ordered given $\mathbf{U}$. Junker and Ellis (1997) characterize such X variables in more generic terms such as conditional association and vanishing conditional dependence. We will call expression (6) the condition of positive association between the sets, $\{X_1, \cdots, X_I\}$ and $\{U_1, \cdots, U_K\}$.

3 Similarity Between BN Models

Consider two BN models which are the same in model structure but not necessarily in distribution. Let the graph be denoted by $\mathcal{G} = (V,\ E)$ where V and E are, respectively, the set of nodes and the set of arrows between nodes. If two nodes u and v are connected by an arrow from u to v, the arrow is represented by (u, v). So, (v, u) cannot be in E if $(u, v) \in E$. Such a graph is called a directed acyclic graph (DAG) when there is no sequence of arrows, $(u_i, u_{i+1}),\ i = 1, 2, \cdots, n$, such that $u_1 = u_{n+1}$. As for a DAG $\mathcal{G}$, we define a set $pa(v)$ for a node v, as $pa(v) = \{u;\ (u, v) \in E\}$.

For two BN models, BN_1 and BN_2, of $\mathbf{X}_V$, assume that they have the same graph, say $\mathcal{G}$, and denote their probability distributions by $\mathcal{D}_1$ and $\widehat{\mathcal{D}_1}$ respectively. In other words, the BN models, BN_i, $i = 1, 2$, are defined as $BN_i = (\mathcal{G}, D_i)$. The probability function P^m of $\mathcal{D}_m$, $m = 1, 2$, is expressed as

$$P^m(x_V) = \prod_{v \in V} P^m_{v|pa(v)}(x_v|\mathbf{x}_{pa(v)}).$$

In other words, $\mathcal{D}_m$ is defined in terms of the conditional distributions of X_v conditional on its parent variables. In this respect, we can compare the two distributions by comparing the conditional or marginal distributions of each variable between the two models. We define the similarity measure between BN_1 and BN_2 as

$$\sum_{v\in V}\sum_{x_v}\sum_{\mathbf{x}_{pa(v)}}\left(P^1_{v|pa(v)}(x_v|x_{pa(v)}) - P^2_{v|pa(v)}(x_v|\mathbf{x}_{pa(v)})\right)^2.$$

We denote the support of X_v by $\mathcal{X}_v$ and let, for $A \subseteq V$, $\mathcal{X}_A = \prod_{v\in A}\mathcal{X}_v$. If all the variables are binary, the similarity measure can be re-expressed as

$$\sum_{v\in V}\sum_{\mathbf{x}\in\mathcal{X}_{pa(v)}}\left(P^1_{v|pa(v)}(1|\mathbf{x}) - P^2_{v|pa(v)}(1|\mathbf{x})\right)^2. \tag{7}$$

We will denote this measure by $\psi(BN_1,\ BN_2)$.

We can think of a probability distribution for $P^1_{v|pa(v)}(1|\mathbf{y}),\ \mathbf{y}\in\mathcal{X}_{pa(v)}$, regarding $P^1_{v|pa(v)}(1|\mathbf{y})$ itself as a random variable, under the condition that the positive association condition (6) is satisfied among $X_v,\ v\in V$. We will denote the distribution of $P^1_{v|pa(v)}(1|y),\ y\in\mathcal{X}_{pa(v)}$, by $\mathcal{D}^v_1$.

Now suppose that $\widehat{\mathcal{D}_1}$ is not known and that we want to find a distribution, say $\mathcal{D}_1$, for which $E(\psi(BN_1,\ BN_2))$ is minimized where the expectation is made with respect to the distribution $\mathcal{D}_1$. By the definition of the similarity measure in (7), it follows that $E(\psi(BN_1,\ BN_2))$ is minimized when $\mathcal{D}_1$ consists of the conditional probabilities of $X_v,\ v\in V$ conditional on $\mathbf{X}_{pa(v)}$ which are given by

$$E(P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)}))$$

where the expectation is made with respect to $\mathcal{D}_1$.

For X_v, we denote by $\delta_v(\mathbf{x})$ the number of 1's in the vector of $\mathbf{x}_{pa(v)}$ and by $\kappa_v(\mathbf{x})$ the number of vectors $\mathbf{y}\in\mathcal{X}_{pa(v)}$ such that $\delta_v(\mathbf{y}) = \delta_v(\mathbf{x})$.

Theorem 3. *Let the expression in (7) be modified as*

$$Q = \sum_{v\in V}\sum_{\mathbf{x}\in\mathcal{X}_{pa(v)}}\left(P^1_{v|pa(v)}(1|\mathbf{x}) - g_v(\delta_v(\mathbf{x}))\right)^2$$

for some functions $g_v,\ v\in V$, where g_v is defined on the set $\{0,1,\cdots,|pa(v)|\}$. If we regard $P^1_{v|pa(v)}(1|\mathbf{x})$ as random variables with distribution $\mathcal{D}_1$, then Q is minimized when

$$g_v(\delta_v(\mathbf{x})) = \frac{1}{\kappa_v(\mathbf{x})}\sum_{\substack{\mathbf{y}\in\mathcal{X}_{pa(v)}:\\ \delta_v(\mathbf{y})=\delta_v(\mathbf{x})}} E(P^1_{v|pa(v)}(1|\mathbf{y})). \tag{8}$$

Proof: Since Q is quadratic in $g_v(\delta(\mathbf{x}))$, we have

$$\begin{aligned}\frac{dQ}{dg_v(\delta_v(\mathbf{x}))} &= \frac{d}{dg_v(\delta_v(\mathbf{x}))}\sum_v \sum_{\mathbf{y}\in\mathcal{X}_{pa(v)}} E\left(P^1_{v|pa(v)}(1|\mathbf{x}) - g_v(\delta_v(\mathbf{x}))\right)^2 \\ &= 2 \sum_{\substack{\mathbf{y}\in\mathcal{X}_{pa(v)}:\\ \delta_v(\mathbf{y})=\delta_v(\mathbf{x})}} E\left(P^1_{v|pa(v)}(1|\mathbf{y}) - g_v(\delta_v(\mathbf{x}))\right) = 0.\end{aligned}$$

Thus, Q is minimized when

$$g_v(\delta_v(\mathbf{x})) = \frac{1}{\kappa_v(\mathbf{x})} \sum_{\substack{\mathbf{y}\in\mathcal{X}_{pa(v)}:\\ \delta_v(\mathbf{y})=\delta_v(\mathbf{x})}} E\left(P^1_{v|pa(v)}(1|\mathbf{y})\right).$$

□

Suppose that $|pa(v)| = 3$. Then

$$\mathcal{X}_{pa(v)} = \{\underbrace{(0,0,0)}_{0}, \underbrace{(0,0,1),(0,1,0),(1,0,0)}_{1}, \underbrace{(0,1,1),(1,0,1),(1,1,0)}_{2}, \underbrace{(1,1,1)}_{3}\}, \quad (9)$$

where the elements ($\mathbf{x}$'s) are grouped according to $\delta(\mathbf{x})$. For four vectors, $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$, in (9), with $\delta(\mathbf{x}^i) = i$, we have that

$$\mathbf{x}^0 \prec \mathbf{x}^i \prec \mathbf{x}^3,\ i = 1, 2. \quad (10)$$

But not every pair in $\mathcal{X}_{pa(v)}$ is ordered. For instance, $(0,1,0)$ and $(1,0,1)$ are not ordered.

Recall that, under the positive association condition, $P_{v|pa(v)}(1|\mathbf{x}) \le P_{v|pa(v)}(1|\mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{pa(v)}$, when $\mathbf{x} \preceq \mathbf{y}$. In constructing a BN model for $\mathbf{X}_V$ under the positive association condition on $P_{v|pa(v)}$, we can think of assigning probability values as follows:

For simplicity of argument, we consider the case that $|pa(v)| = 3$. If we are given 8 values between 0 and 1 in a random manner from some distribution, we assign these values to $P_{v|pa(v)}(1|\mathbf{x}),\ \mathbf{x} \in \mathcal{X}_{pa(v)}$, in the order of the elements in (9) from small to large allowing order distortions between some of $P_{v|pa(v)}(1|\mathbf{x})$ values for which the $\mathbf{x}$'s are not comparable in the sense of $\preceq$.

Since the values are assigned in a random manner to the $P_{v|pa(v)}(1|\mathbf{x})$'s for which the $\mathbf{x}$'s are not comparable with each other, Theorem 3 says that, if the conditional probability of X_v conditional on $\mathbf{X}_{pa(v)}$ depends upon $\delta_v(\mathbf{X})$ only in a BN model and minimizes $E(Q)$, then the desired conditional probability is given by the right-hand side of (8).

Suppose we are given a BN model $BN_1 = (\mathcal{G}, \mathcal{D}_1)$. A BN model for which g_v values are used in place of $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$ will be called the similar BN model of BN_1 and be denoted by $\widehat{BN}_1$. A BN model BN_1 itself is not ready for computation for classification since the conditional probability $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$ is yet a random variable therein. When the conditional probability is replaced with a list of numeric values corresponding to the values of $\mathbf{x}_{pa(v)}$ for every node v, we will call the BN model an actual BN model of BN_1. In other words, a BN_1 is to its actual BN model what a random variable is to its outcome or observed value.

We will compare the performance of $\widehat{BN_1}$ with the actual BN model of BN_1 in next section.

4 An Example of Similar BN Models

The agreement level of the classifications between the two models, an actual model of BN_1 and $\widehat{BN_1}$, is defined as follows:

> Suppose we compare predictions for subject j with regard to X_k and denote the class levels from an actual BN model and the similar BN model of BN_1, respectively, by Y^a_{jk} and Y^s_{jk}. We define $e_{jk} = Y^s_{jk} - Y^a_{jk}$. So, if the class levels range from 1 through L, then $-L+1 \le e_{jk} \le L-1$. Suppose that there are N subjects for whom classifications are to be made. Then we can define the agreement level up to difference d ($d > 0$) by
>
> $$\alpha^k_d = \sum_{h:\ |h| \le d} r_{kh} \tag{11}$$
>
> where
>
> $$r_{kh} = \frac{\text{the number of cases that } e_{jk} = h}{N}.$$

We will use α^k_0 and α^k_1 as the two primary indices of agreement. As for BN_1, we will consider the beta distribution with parameters a and b (denoted by $\mathcal{B}(a,b)$) and some of its variations for $P^1_{v|pa(v)}(1|\mathbf{x}),\ \mathbf{x} \in \mathcal{X}_{pa(v)}$. Since $P^1_{v|pa(v)}(1|\mathbf{x})$ are ordered according to $\mathbf{x}$ under the positive association condition, we may use a set of order statistics of a random sample from $\mathcal{B}(a,b)$ for $P^1_{v|pa(v)}(1|\mathbf{x})$.

Let V be partitioned into $\{V_1, V_2\}$ and suppose that we are interested in classifying for $\mathbf{X}_{V_2}$ given an outcome of $\mathbf{X}_{V_1}$. In Figure 2, $\mathbf{X}_{V_2} = (U_1, \cdots, U_6)$. In the rest of this section, we obtain the α values for $\mathbf{X}_v,\ v \in V_2$, by a simulation experiment for a given distribution $\mathcal{D}_1$. Denote by $\mathbf{x}_i$ the outcome of $\mathbf{X}$ for the i^{th} case of data and by $\mathbf{x}_{ik}$ the k^{th} component of $\mathbf{x}_i$. For a set A, let $\mathbf{x}_{i,A} = (x_{ik},\ k \in A)$.

In the simulation, we obtained the α values as follows:

(i) For a given $BN_1 = (\mathcal{G}, \mathcal{D}_1)$, construct a similar BN model, $\widehat{BN_1}$.

(ii) Generate a set of numeric values of the conditional probabilities, $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$ from $\mathcal{D}_1$ and construct an actual BN model using the generated values.

(iii) Generate N vector values, $\mathbf{x}_i,\ i = 1, \cdots, N$, of $\mathbf{X}_V$ from the actual model.

(iv) For each $k \in V_2$, obtain Y^a_{ik} and Y^s_{ik} for $i = 1, 2, \cdots, N$. $Y^a_{ik} = h$ if the rank of $P(X_{ik} = 1|\mathbf{x}_{i,V_1})$ is larger than $N(h-1)/L$ and not larger than Nh/L, where $P(X_{ik} = 1|\mathbf{x}_{i,V_1})$'s are computed in the actual model and the rank is assigned in the reverse order of the magnitude of $P(X_{ik} = 1|\mathbf{x}_{i,V_1})$, $i = 1, 2, \cdots, N$. Y^s_{ik}'s are obtained analogously except that $P(X_{ik} = 1|\mathbf{x}_{i,V_1})$'s are computed in $\widehat{BN_1}$.

(v) For each $k \in V_2$, obtain r_{kh}, $-L+1 \le h \le L-1$.

(vi) Repeat steps (ii)-(v) B times and compute the average of the B r_{kh} values for each $k \in V_2$.

It is important to note in the above process that the values of $\mathbf{X}_V$ are generated from an actual model of BN_1 and the comparison of class levels between the actual

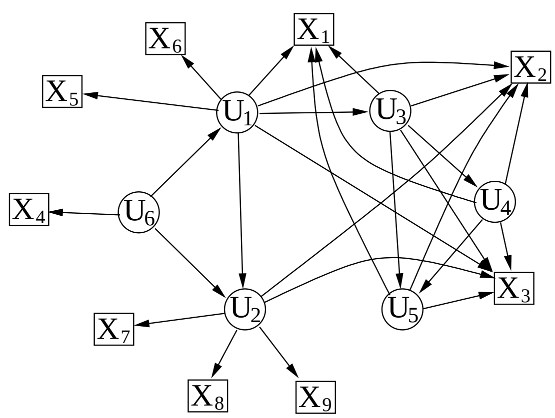


Figure 2: A BN where classifications are made for U variables based on X variables.

and the similar models are based on the conditional probabilities, $P(X_{ik} = 1|\mathbf{x}_{i,V_1})$, $k \in V_2$. We denote by $\bar{r}_{kh}$ the average of the r_{kh} values and replace the r_{kh} in (11) by $\bar{r}_{kh}$ as in

$$\alpha_d^k = \sum_{h:\ |h| \leq d} \bar{r}_{kh}. \tag{12}$$

Since B different sets of numeric values of the conditional probabilities, $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$, are used for actual BN models from BN_1, the α values as in (12) can be regarded as an estimate of the mean of the agreement level of the classifications between the actual BN model and the similar BN model, when $\mathcal{D}_1$ is the true distribution for $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$. In the simulation experiment below, we set $N = 1,000$ and $B = 100$.

We will take as an example $\mathcal{B}(0.2, 0.4)$ for $\mathcal{D}_1$, $v \in V$. When the conditional probabilities, $\{P^1_{v|pa(v)}(1|\mathbf{x}),\ v \in V\}$, are given as order statistics from $\mathcal{U}(0,1) = \mathcal{B}(1,1)$, the $P^1_{v|pa(v)}$ values tend to be evenly dispersed over the interval between 0 and 1. But the $P^1_{v|pa(v)}(1|\mathbf{x})$ values are more likely to appear near the end points of the interval when they are from $\mathcal{B}(0.2, 0.4)$. In the case of educational testing, if X_v is an item score (1 for correct answer, 0 otherwise) and $\mathcal{X}_{pa(v)}$ is a set of vectors of the knowledge states (1 for good enough, 0 otherwise) of the knowledge units that are required for the test item, then the $P^1_{v|pa(v)}$ values tend to be small when $\delta(\mathbf{x}_{pa(v)}) < |pa(v)|$ and near 1 otherwise (see, for example, Mislevy(1994)). In this respect, the Beta distribution is a reasonable choice for $\mathcal{D}_1$.

The mean of the jth smallest among a random sample of size n from a Beta distribution is not available in a closed form, and the g_v values are computed by a numerical approach. Table 1 lists the g_v values for $|pa(v)| \leq 6$, which are used for constructing $\widehat{BN_1}$ corresponding to $\mathcal{D}_1 = \mathcal{B}(0.2, 0.4)$.

It is indicated in Table 2 when $L = 5$ that, on average, the classifications from the two types of models, the actual and the similar models, exactly agree for about 57% of the cases (see the column of α_0 in the table) and the average of the six values in the last column is about 0.93.

It is worthwhile to note in Table 1 that the $g_v(0)$ values become very small as $|pa(v)|$ increases. If we consider educational testing with multiple choice tests, these values are unreasonably small since multiple choice tests allow guessing for selecting response options. An example of the distribution $\mathcal{D}_1$ for this situation is

$$0.85 \cdot \mathcal{B}(a,\ b) + 0.1, \tag{13}$$

Table 1: The g_v values of (8) for $|pa(v)| \leq 6$ when $\mathcal{D}_1$ is $\mathcal{B}(0.2, 0.4)$

$$|pa(v)| = 0: \quad g_v(0) = 0.3333$$

$$|pa(v)| = 1: \quad g_v(i) = \begin{cases} 0.1328 & \text{if } i = 0 \\ 0.5338 & \text{if } i = 1 \end{cases}$$

$$|pa(v)| = 2: \quad g_v(i) = \begin{cases} 0.0299 & \text{if } i = 0 \\ 0.2783 & \text{if } i = 1 \\ 0.7567 & \text{if } i = 2 \end{cases}$$

$$|pa(v)| = 3: \quad g_v(i) = \begin{cases} 0.0035 & \text{if } i = 0 \\ 0.0801 & \text{if } i = 1 \\ 0.5072 & \text{if } i = 2 \\ 0.9012 & \text{if } i = 3 \end{cases}$$

$$|pa(v)| = 4: \quad g_v(i) = \begin{cases} 0.0002 & \text{if } i = 0 \\ 0.0116 & \text{if } i = 1 \\ 0.2185 & \text{if } i = 2 \\ 0.7508 & \text{if } i = 3 \\ 0.9724 & \text{if } i = 4 \end{cases}$$

$$|pa(v)| = 5: \quad g_v(i) = \begin{cases} 0.0000 & \text{if } i = 0 \\ 0.0010 & \text{if } i = 1 \\ 0.0537 & \text{if } i = 2 \\ 0.4580 & \text{if } i = 3 \\ 0.9100 & \text{if } i = 4 \\ 0.9940 & \text{if } i = 5 \end{cases}$$

$$|pa(v)| = 6: \quad g_v(i) = \begin{cases} 0.0000 & \text{if } i = 0 \\ 0.0000 & \text{if } i = 1 \\ 0.0082 & \text{if } i = 2 \\ 0.1829 & \text{if } i = 3 \\ 0.7136 & \text{if } i = 4 \\ 0.9749 & \text{if } i = 5 \\ 0.9988 & \text{if } i = 6 \end{cases}$$

which means that guessing shrinks the range of the conditional probabilities from $\mathcal{B}(a, b)$ to $0.85 \cdot \mathcal{B}(a, b) + 0.1$.

Given the g_v values of (8) corresponding to the $\mathcal{D}_1 = \mathcal{B}(0.2, 0.4)$, we can easily obtain the g_v values corresponding to $\mathcal{D}_1 = a\mathcal{B}(0.2, 0.4) + (0.95 - a)$. In expression (8), $P^1_{v|pa(v)}(1|\mathbf{y})$ is regarded as a random variable from a distribution $\mathcal{D}_1$. So if we denote by g_v^a the g_v values corresponding to $\mathcal{D}_1 = a\mathcal{B}(0.2, 0.4) + (0.95 - a)$, and by g_v^* corresponding to $\mathcal{D}_1 = \mathcal{B}(0.2, 0.4)$, then it follows from expression (8) that

$$g_v^a = a g_v^* + (0.95 - a). \tag{14}$$

Table 2: α and $\bar{r}_{kh}$ values with $L = 5$ and $B = 50$ for the model in Figure 2 when $\mathcal{D}_1$ is $\mathcal{B}(0.2, 0.4)$

U_i	$\bar{r}_{k,-4}$	$\bar{r}_{k,-3}$	$\bar{r}_{k,-2}$	$\bar{r}_{k,-1}$	α_0	$\bar{r}_{k1}$	$\bar{r}_{k2}$	$\bar{r}_{k3}$	$\bar{r}_{k4}$	α_1
U_1	0.000	0.002	0.023	0.186	0.544	0.182	0.050	0.012	0.001	0.912
	(0.000)	(0.007)	(0.025)	(0.056)	(0.100)	(0.053)	(0.050)	(0.031)	(0.004)	(0.083)
U_2	0.000	0.001	0.024	0.190	0.565	0.180	0.035	0.006	0.001	0.934
	(0.000)	(0.002)	(0.026)	(0.060)	(0.097)	(0.051)	(0.030)	(0.016)	(0.003)	(0.056)
U_3	0.000	0.002	0.017	0.173	0.611	0.166	0.026	0.004	0.000	0.950
	(0.000)	(0.008)	(0.026)	(0.057)	(0.138)	(0.061)	(0.035)	(0.009)	(0.001)	(0.058)
U_4	0.001	0.001	0.021	0.165	0.607	0.179	0.026	0.001	0.000	0.951
	(0.004)	(0.005)	(0.025)	(0.059)	(0.113)	(0.066)	(0.037)	(0.003)	(0.001)	(0.054)
U_5	0.000	0.002	0.020	0.159	0.624	0.173	0.020	0.002	0.000	0.957
	(0.000)	(0.005)	(0.022)	(0.071)	(0.117)	(0.058)	(0.026)	(0.005)	(0.000)	(0.043)
U_6	0.000	0.005	0.045	0.193	0.486	0.180	0.061	0.020	0.009	0.860
	(0.001)	(0.011)	(0.040)	(0.075)	(0.126)	(0.055)	(0.050)	(0.048)	(0.037)	(0.127)
Average	0.000	0.002	0.025	0.178	**0.573**	0.177	0.036	0.007	0.002	**0.927**

NOTE: The values in the parentheses are standard deviations.

We can see in Table A.1 in Appendix that the α_0 and α_1 values when $\mathcal{D}_1 = 0.85\mathcal{B}(0.2, 0.4) + 0.1$ are about 60% and 95%, respectively. In both of the cases, the α_0 and α_1 values are larger than 50% and 90%, respectively. We will extend the simulation experiment to a more general version of $\mathcal{D}_1$ in next section.

5 Agreement Levels Between Two Types of BN Models

In the preceding section we compared classifications between an actual model of BN_1 and its similar model, $\widehat{BN_1}$, where $\mathcal{D}_1 = \mathcal{B}(0.2, 0.4)$. In this section we will consider $\mathcal{D}_1$ as a mixed distribution which is given by

$$C + (0.95 - C) \times \mathcal{B}(A, B) \tag{15}$$

where A, B, C are independent Uniform random variables over the interval (0.1, 0.95). This means that $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$ as appearing in Theorem 3 has its expected value $E(E(P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})|A, B, C)) = 0.7375$, where we note that the mean of the Beta random variable with the random parameters A and B is equal to 0.5. We selected the Uniform distribution over the interval (0.1, 0.95) to avoid too small or too large values of $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$.

The mixed distribution is equivalent to the family of distributions,

$$\mathcal{F} = \{c + (0.95 - c) \times \mathcal{B}(a, b);\ \ a, b, c \in [0.1, 0.95]\} \tag{16}$$

where each distribution in $\mathcal{F}$ is equally likely. Suppose that, for each node v in a BN model, its conditional probability $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$ follows a distribution in $\mathcal{F}$ and that the distributions may be different across the nodes of the BN model. Also suppose that we have an actual BN model of BN_1 with $\mathcal{D}_1$ given by (15). Its similar BN model is constructed, in light of Theorem 3, by using g_v's which are obtained by taking the distribution of $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$ as

$$0.525 + 0.425 \times \mathcal{B}(0.525, 0.525), \tag{17}$$

which we will denote by $\widehat{\mathcal{D}_1}$.

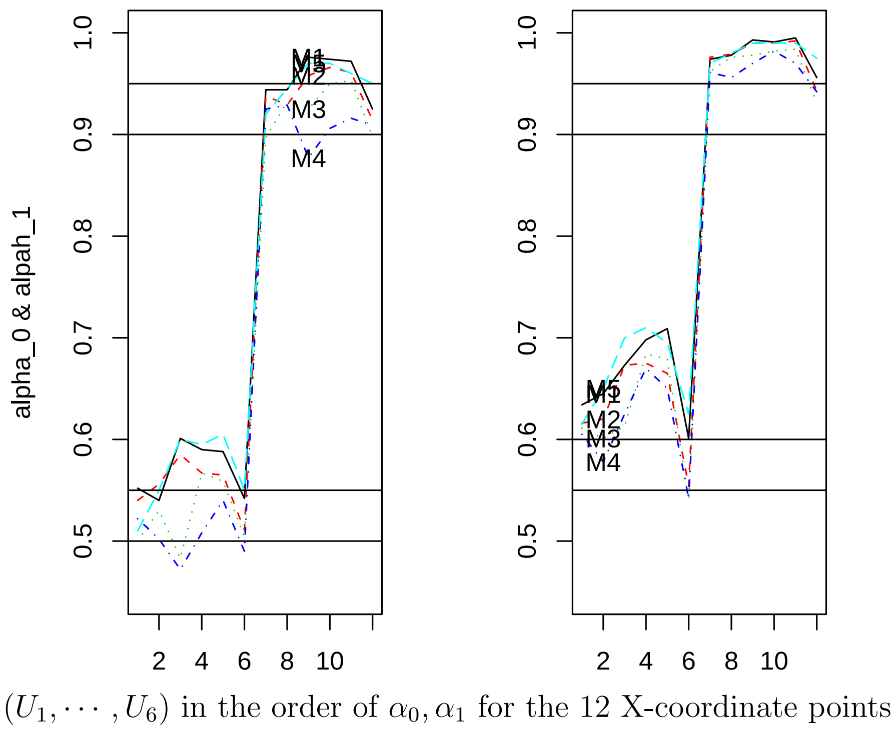
We computed the agreement levels between the two types of BN models, BN_1 and $\widehat{BN_1}$, and summarized the levels in terms of the three quantities, the first(Q_1), the second(Q_2), and the third(Q_3) quartiles of B(=100) (α_0, α_1)-values which are obtained by iterating the comparison of the classifications between the two models 100 times. At each iteration, we construct an actual model of BN_1 and use the g_v's

Table 3: α_0 and α_1 values for the model in Figure 2.
(a) $\mathcal{D}_1$ as given by (15) is used for actual models and $\widehat{\mathcal{D}_1}$ as in (17) for the similar model.

	α_0						α_1					
	U_1	U_2	U_3	U_4	U_5	U_6	U_1	U_2	U_3	U_4	U_5	U_6
min	0.316	0.305	0.392	0.270	0.281	0.251	0.725	0.695	0.662	0.740	0.760	0.564
Q_1	0.466	0.464	0.522	0.508	0.512	0.446	0.894	0.897	0.936	0.912	0.929	0.837
Q_2	0.552	0.540	0.601	0.590	0.588	0.542	0.944	0.944	0.976	0.974	0.972	0.925
Q_3	0.655	0.674	0.701	0.691	0.691	0.626	0.979	0.979	0.992	0.992	0.993	0.975
max	0.853	0.875	0.824	0.862	0.954	0.874	1.000	1.000	1.000	1.000	1.000	1.000

(b) $\widehat{\mathcal{D}_1}$ as in (17) is used for both the actual and the similar models.

	α_0						α_1					
	U_1	U_2	U_3	U_4	U_5	U_6	U_1	U_2	U_3	U_4	U_5	U_6
min	0.341	0.324	0.328	0.350	0.376	0.251	0.825	0.746	0.850	0.751	0.803	0.600
Q_1	0.548	0.514	0.589	0.593	0.625	0.501	0.938	0.937	0.969	0.962	0.980	0.893
Q_2	0.634	0.645	0.673	0.698	0.709	0.601	0.974	0.978	0.993	0.991	0.995	0.956
Q_3	0.731	0.714	0.751	0.779	0.797	0.675	0.993	0.994	1.000	1.000	1.000	0.988
max	0.891	0.868	0.893	0.926	0.921	0.864	1.000	1.000	1.000	1.000	1.000	1.000



(a) $\mathcal{D}_1$ for the actual model and $\widehat{\mathcal{D}_1}$ for the similar model.

(b) $\widehat{\mathcal{D}_1}$ for both of the actual and the similar models.

Figure 3: Medians (Q_2) of the α values for the six U variables which are listed in Tables 3 and A.2. The first six of the twelve points on the X-axis indicate the six U variables, $U_1, \cdots, U_6$, for α_0, and the remaining six for α_1. M1, $\cdots$, M5 in the graph are another notation of $M_1, \cdots, M_5$, respectively.

from $\widehat{\mathcal{D}_1}$ in expression (17) for the similar model, $\widehat{BN_1}$. The result is summarized in Table 3(a) for the BN model in Figure 2.

The mixed distribution $\mathcal{D}_1$ can be regarded as representing our uncertainty on the true values of the conditional probabilities $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$; while $\widehat{\mathcal{D}_1}$ as representing the mean of the unknown true values. In Table 3(b), we summarize the comparison result between the actual and similar models which share the same distribution $\widehat{\mathcal{D}_1}$. In other words, the comparison of the classifications is made between two models for which $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$'s are obtained from the same distribution $\widehat{\mathcal{D}_1}$. The comparison thus shows us the agreement level which is affected by the chancy fluctuation of the rank-orders between the two models.

For convenience, we will call the four BN models in Figures 2, A.1, $\cdots$, A.4, models 1, 2, $\cdots$, 5, respectively, in the order of their figures, where the last four figures are in Appendix. The five BN models are different as follows. Model 2 is obtained by adding to model 1 six arrows between U variables and X variables and one arrow between U variables. Model 3 is obtained by removing arrows from model 2 so that U_1, U_4, and U_6 become root nodes and the three sets, $\{U_1, U_2, U_3, U_5\}, \{U_4\}$, and $\{U_6\}$, are marginally independent among themselves. By removing all the arrows between the U variables in model 3 we obtain model 4 so that all the U variables are marginally independent among themselves. On the contrary, model 5 is obtained from model 2 by connecting every pair of the U variables with an arrow.

The main difference between the first two of the five BN models is that all the X variables each has multiple parent nodes while it is not the case in model 1. As for models 2, 3, 4, and 5, we see more independencies among the U variables in the order of models 5, 2, 3, 4. We considered this variation among the model structures to see if this variation affects the agreement level, which is not likely as is shown in Figure 3. The Q_2 values (or medians) of the α values as listed in Tables 3 and A.2 are depicted in the figure. The first six points on each curve are the Q_2 values of the α_0 values for $U_1, \cdots, U_6$, and the remaining six points correspond to the Q_2 values of the α_1 values. We can see in both of the panels in Figure 3 that models M_1 (or M_5) and M_4 are farthest apart on average among all the possible pairs of the models. However, we can check in Tables 3 and A.2 that the medians of the α values from model 1 belong to the mid-50% interval of the corresponding α values from model 4 or vice versa for at least half of the U variables. This indicates that the model structure or marginal independencies among U variables may not be very influential on the agreement level between the two types of models, the actual and the similar models.

As aforementioned, $\mathcal{D}_1$ is a mixed distribution in Table 3(a). We thus can say that $\mathcal{D}_1$ is to $\widehat{\mathcal{D}_1}$ what a random variable is to its mean. When we are not sure of the distribution of $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$, we can use an 'averaged' version of the distribution in the form of $\widehat{\mathcal{D}_1}$. Comparison of the agreement levels between panels (a) and (b) are important in this context. It is worthwhile to note in the table that, for every variable U_i, Q_1 of the α values in panel (b) lies between Q_1 and Q_3 in panel (a). The same result holds for other models in Figures A.1, A.2, A.3, and A.4, as is summarized in Table A.2 in Appendix. This result suggests that when we are not sure of the true values of

$P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$, we may use instead the 'averaged' version $\widehat{\mathcal{D}_1}$ for classification and use the agreement level of the classifications between two types of models, BN_1 and $\widehat{BN_1}$, as an index of credibility in classifying via an alternative model $\widehat{BN_1}$.

6 Concluding Remarks

When a BN model contains unobservable variables, we usually use an EM method Dempster, Laird, and Rubin (1977) for estimating the conditional probabilities of the variables in the BN model. If many variables are involved in the BN model, then the parameter estimation is time-consuming and the estimates tend to be contaminated with larger errors than when the model is of a moderate size (Kim 2002). In this respect, when the BN model is of a large size and involves unobservable binary variables and the classification is given in terms of rank-order based class levels of probabilities, the similar BN model is recommended as a reasonable alternative.

The notion of similarity between models is defined under the premise that the models are of the same model structure. In the definition, we assume that the underlying distribution, $\mathcal{D}_1$, is known. But, in reality, we have data for observable variables only and the classification is usually made for unobservable variables. It may be desirable, when a timely service of classification is required and an acceptable development of a BN model is yet in progress, that we use some alternative model for the classification which is made within some pre-specified tolerable limits. The simulation experiment strongly suggests, under the positive association condition, that we may use a similar model, $\widehat{BN_1}$, for classification in addition to the agreement levels which are computed by comparing the two types of models, BN_1 and $\widehat{BN_1}$.

A main idea behind the proposed similar model approach for classification is analogous to Bayesian approach for statistical inference. We use a $\widehat{\mathcal{D}_1}$ distribution for the similar model which is obtained under a prior assumption on the conditional probabilities, $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$. Suppose that we have a data set of the observable variables, $\mathbf{X}_{V_1}$, and that we are not sure of the probability models of $\mathbf{X}_V$ except its model structure, which is given in a BN. Then we assume a set of prior distributions for the parameters of the distribution of the conditional probabilities, $P^1_{v|pa(v)}(1|\mathbf{x}_{pa(v)})$, which all together form a mixed distribution, $\mathcal{D}_1$, and regard the data as obtained from a distribution in $\mathcal{F}$ in expression (16). We use the similar model with $\widehat{\mathcal{D}_1}$ for classification along with the agreement level which is obtained in the same way as for

Table 3(a).

It is pointed out in Simon (1981) that discrete features are closer to a knowledge-level representation than continuous ones, and we often express our opinions or feelings using the rank-order based categorized terms such as 'very little,' 'little,' 'medium,' 'much,' and 'very much.' In particular, for survey and assessment in medicine, education, and marketing among others, we usually use category levels instead of continuous values. The similar model as proposed in this paper can be used as a surrogate classifier of a group of subjects when consistency of classification is required, provided that we can reasonably assume a mixed distribution $\mathcal{D}_1$ as illustrated in expression (15).

Appendix

Table A.1: α and $\bar{r}_{kh}$ values with $L = 5$ and $B = 50$ for the model in Figure 2 when $\mathcal{D}_1$ is $0.85 \cdot \mathcal{B}(0.2, 0.4) + 0.1$

U_i	$\bar{r}_{k,-4}$	$\bar{r}_{k,-3}$	$\bar{r}_{k,-2}$	$\bar{r}_{k,-1}$	α_0	$\bar{r}_{k1}$	$\bar{r}_{k2}$	$\bar{r}_{k3}$	$\bar{r}_{k4}$	α_1
U_1	0.000	0.004	0.029	0.178	0.594	0.176	0.018	0.001	0.000	0.948
	(0.001)	(0.008)	(0.024)	(0.042)	(0.110)	(0.048)	(0.025)	(0.003)	(0.001)	(0.051)
U_2	0.000	0.003	0.030	0.172	0.602	0.172	0.020	0.001	0.000	0.947
	(0.000)	(0.004)	(0.029)	(0.040)	(0.111)	(0.053)	(0.028)	(0.003)	(0.000)	(0.057)
U_3	0.000	0.003	0.029	0.190	0.564	0.189	0.025	0.000	0.000	0.942
	(0.001)	(0.008)	(0.029)	(0.046)	(0.117)	(0.062)	(0.038)	(0.001)	(0.000)	(0.066)
U_4	0.000	0.002	0.022	0.163	0.635	0.168	0.010	0.000	0.000	0.966
	(0.000)	(0.004)	(0.027)	(0.046)	(0.121)	(0.066)	(0.016)	(0.000)	(0.000)	(0.038)
U_5	0.000	0.002	0.021	0.162	0.639	0.165	0.011	0.001	0.000	0.966
	(0.002)	(0.005)	(0.024)	(0.053)	(0.127)	(0.071)	(0.015)	(0.004)	(0.000)	(0.034)
U_6	0.002	0.007	0.040	0.185	0.550	0.171	0.036	0.007	0.002	0.906
	(0.007)	(0.013)	(0.040)	(0.052)	(0.138)	(0.059)	(0.039)	(0.015)	(0.006)	(0.092)
Average	0.000	0.004	0.029	0.175	**0.597**	0.173	0.020	0.002	0.000	**0.946**

NOTE: The values in the parentheses are standard deviations.

Three more results in addition to the result of section 5 follow. These results correspond to three different BN models as displayed in Figures A.1, A.2, A.3, and A.4 below.

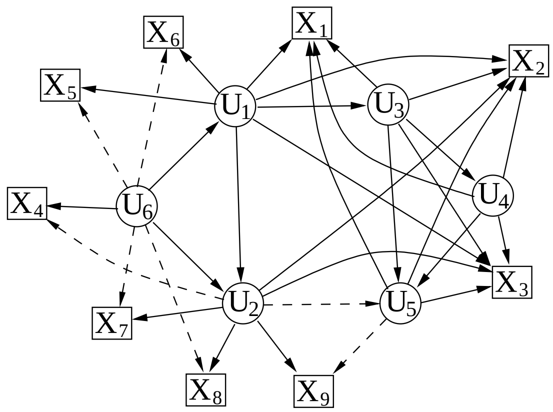


Figure A.1: A BN which is obtained by adding dashed arrows to the BN in Figure 2.

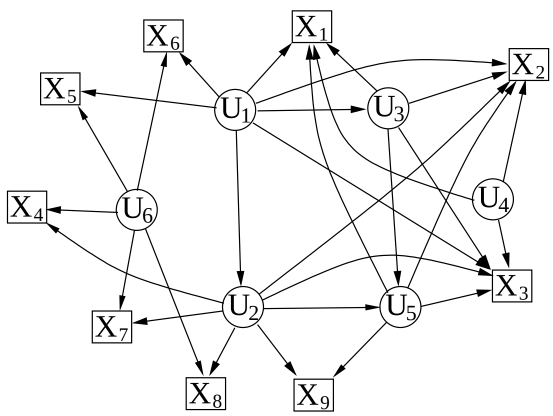


Figure A.2: A BN which is obtained by removing some arrows from the BN in Figure A.1.

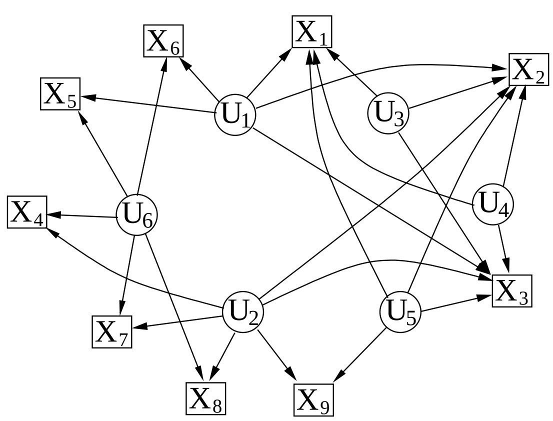


Figure A.3: A BN which is obtained by removing all the arrows between the U nodes from the BN in Figure A.2.

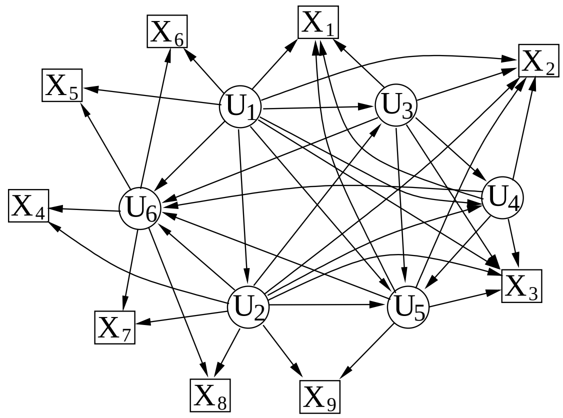


Figure A.4: A BN which is obtained by connecting all the U nodes in the order of node-labels in the BN in Figure A.2.

Table A.2: α_0 and α_1 values for the models in Figures A.1, A.2, A.3, and 3.

For model 2 in Figure A.1:

(a) $\mathcal{D}_1$ given by (15) is used for actual models and $\widehat{\mathcal{D}_1}$ in (17) for the similar model.

	α_0						α_1					
	U_1	U_2	U_3	U_4	U_5	U_6	U_1	U_2	U_3	U_4	U_5	U_6
min	0.351	0.332	0.267	0.335	0.323	0.232	0.708	0.715	0.633	0.806	0.779	0.661
Q_1	0.480	0.470	0.481	0.497	0.469	0.437	0.907	0.895	0.909	0.930	0.932	0.864
Q_2	0.540	0.556	0.585	0.567	0.565	0.511	0.938	0.929	0.958	0.966	0.961	0.914
Q_3	0.609	0.605	0.685	0.686	0.647	0.582	0.963	0.967	0.991	0.990	0.988	0.951
max	0.780	0.832	0.854	0.928	0.907	0.725	0.999	1.000	1.000	1.000	1.000	0.997

(b) $\widehat{\mathcal{D}_1}$ as in (17) is used for both the actual and the similar models.

	U_1	U_2	U_3	U_4	U_5	U_6	U_1	U_2	U_3	U_4	U_5	U_6
min	0.263	0.366	0.384	0.339	0.318	0.286	0.808	0.832	0.892	0.888	0.856	0.587
Q_1	0.527	0.550	0.602	0.584	0.585	0.481	0.952	0.949	0.965	0.974	0.971	0.905
Q_2	0.616	0.620	0.673	0.675	0.665	0.554	0.976	0.979	0.990	0.990	0.992	0.943
Q_3	0.687	0.691	0.761	0.764	0.751	0.635	0.994	0.992	1.000	1.000	1.000	0.976
max	0.920	0.863	0.919	0.905	0.957	0.783	1.000	1.000	1.000	1.000	1.000	1.000

(Table A.2 continued)

For model 3 in Figure A.2:

(a) $\mathcal{D}_1$ and $\widehat{\mathcal{D}_1}$ are the same as for model 2.

	α_0						α_1					
	U_1	U_2	U_3	U_4	U_5	U_6	U_1	U_2	U_3	U_4	U_5	U_6
min	0.252	0.267	0.206	0.209	0.307	0.287	0.630	0.749	0.384	0.641	0.705	0.638
Q_1	0.439	0.449	0.405	0.450	0.454	0.434	0.857	0.872	0.857	0.872	0.880	0.860
Q_2	0.500	0.530	0.483	0.567	0.559	0.504	0.897	0.932	0.925	0.950	0.953	0.898
Q_3	0.556	0.614	0.575	0.672	0.653	0.565	0.941	0.972	0.968	0.974	0.983	0.935
max	0.730	0.749	0.812	0.920	0.848	0.775	1.000	1.000	1.000	1.000	1.000	0.984

(b) $\widehat{\mathcal{D}_1}$ as in (17) is used for both the actual and the similar models.

min	0.291	0.324	0.347	0.319	0.259	0.219	0.744	0.763	0.699	0.663	0.766	0.739
Q_1	0.534	0.540	0.527	0.573	0.537	0.466	0.929	0.946	0.947	0.946	0.946	0.891
Q_2	0.611	0.601	0.613	0.684	0.679	0.540	0.962	0.976	0.978	0.982	0.984	0.931
Q_3	0.675	0.679	0.705	0.752	0.761	0.593	0.979	0.993	0.994	0.998	0.999	0.962
max	0.807	0.882	0.854	0.901	0.912	0.818	1.000	1.000	1.000	1.000	1.000	1.000

For model 4 in Figure A.3:

(a) $\mathcal{D}_1$ and $\widehat{\mathcal{D}_1}$ are the same as for model 2.

min	0.215	0.147	0.102	0.113	0.159	0.182	0.439	0.714	0.311	0.213	0.472	0.635
Q_1	0.414	0.424	0.313	0.335	0.426	0.413	0.840	0.855	0.723	0.802	0.840	0.853
Q_2	0.522	0.502	0.472	0.508	0.540	0.490	0.925	0.929	0.877	0.906	0.916	0.909
Q_3	0.676	0.624	0.612	0.611	0.639	0.587	0.959	0.955	0.946	0.971	0.962	0.942
max	0.876	0.919	0.863	0.882	0.939	0.917	1.000	0.996	1.000	1.000	0.999	0.995

(b) $\widehat{\mathcal{D}_1}$ as in (17) is used for both the actual and the similar models.

min	0.223	0.196	0.221	0.240	0.126	0.225	0.512	0.215	0.525	0.578	0.431	0.631
Q_1	0.526	0.484	0.506	0.550	0.505	0.462	0.923	0.921	0.919	0.941	0.929	0.889
Q_2	0.605	0.578	0.625	0.670	0.650	0.542	0.961	0.956	0.970	0.982	0.970	0.942
Q_3	0.722	0.686	0.750	0.781	0.764	0.651	0.987	0.978	0.994	0.999	0.987	0.967
max	0.952	0.876	0.965	0.989	0.942	0.963	1.000	1.000	1.000	1.000	1.000	0.999

For model 5 in Figure 3:

(a) $\mathcal{D}_1$ given by (15) is used for actual models and $\widehat{\mathcal{D}_1}$ in (17) for the similar model.

min	0.170	0.320	0.300	0.240	0.300	0.260	0.620	0.800	0.590	0.720	0.670	0.660
Q_1	0.458	0.490	0.518	0.520	0.500	0.467	0.880	0.907	0.930	0.940	0.910	0.900
Q_2	0.510	0.550	0.600	0.595	0.605	0.550	0.920	0.945	0.970	0.970	0.960	0.950
Q_3	0.590	0.630	0.680	0.710	0.702	0.620	0.960	0.970	0.990	0.990	0.990	0.970
max	0.810	0.780	0.810	0.840	0.830	0.880	1.000	1.000	1.000	1.000	1.000	1.000

(b) $\widehat{\mathcal{D}_1}$ as in (17) is used for both the actual and the similar models.

min	0.340	0.350	0.330	0.390	0.190	0.240	0.720	0.840	0.560	0.650	0.650	0.810
Q_1	0.510	0.577	0.590	0.628	0.610	0.550	0.940	0.960	0.970	0.970	0.970	0.940
Q_2	0.615	0.650	0.700	0.710	0.695	0.625	0.970	0.980	0.990	0.990	0.990	0.975
Q_3	0.712	0.720	0.763	0.780	0.760	0.700	0.990	0.990	1.000	1.000	1.000	0.990
max	0.850	0.920	0.930	0.920	0.900	0.870	1.000	1.000	1.000	1.000	1.000	1.000

Refences

ALBATINEH, A.N. NIEWIADOMSKA-BUGAJ, M, and MIHALKO, D. (2006), "On similarity indices and correction for chance agreement," *Journal of Classification* 23, 301-313.

ANDERSEN, S.K., JENSEN, F.V., OLESEN, K.G., and JENSEN, F. (1989). *HUGIN: A shell for building Bayesian belief universes for expert systems* [computer program]. HUGIN Expert Ltd., Aalborg, Denmark.

BRUSCO, M.J. and STEINLEY, D. (2008), "A binary integer program to maximize the agreement between partitions," *Journal of Classification* 25, 185-193.

COHEN, J. (1960). "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.* **20**, 37-46.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B.(1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society B* 39, 1-38.

ERGO [computer program] Noetic Systems Inc., Baltimore, MD, 1991.

FLEISS, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edn. New York: Wiley.

HOLLAND, P.W. and ROSENBAUM, P.R. (1986). "Conditional association and unidimensionality in monotone latent variable models," *The Annals of Statistics* 14(4), 1523-1543.

JENSEN, F.V.(1996). *An Introduction to Bayesian Networks*, New York: Springer-Verlag.

JUNKER, B.W. and ELLIS, J.L. (1997). "A characterization of monotone unidimensional latent variable models," *The Annals of Statistics* 25(3), 1327-1343.

KIM, S.H. (2002), "Calibrated initials for an EM applied to recursive models of categorical variables," *Computational Statistics and Data Analysis*, 40(1), 97-110.

KIM, S.H. (2005), “Stochastic ordering and robustness in classification from a Bayesian network,” *Decision Support Systems* 39, 253-266.

LAURITZEN, S.L. and SPIEGELHALTER, D.J. (1988), “Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems,” *Journal of the Royal Statistical Soc. B*, 50(2), 157-224 .

MESSATFA, H. (1992), “An algorithm to maximize the agreement between partitions,” *Journal of Classification* 9, 5-15.

MSBNx [computer program], http://research.microsoft.com/msbn/, 2001.

MISLEVY, R.J.(1994), “Evidence and inference in educational assessment,” *Psychometrika* 59(4), 439-483.

PEARL, J. (1988), *Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA.: Margan Kaufmann.

PIRES, A.M. and BRANCO, J.A. (1997), “Comparison of multinomial classification rules,” *Journal of Classification*, 14, 137-145.

SIMON, H.A. (1996). *The Sciences of the Artificial.* 3rd ed. Cambridge: The MIT Press.

SPITZER, R.L., COHEN, J., FLEISS, J.L., and ENDICOTT, J. (1967), “Quantification of agreement in psychiatric diagnosis,” *Arch. Gen. Psychiatry* **17**, 83-87.