Stochastic Ordering and Robustness in Classification from a Bayesian Network

Sung-Ho Kim

Division of Applied Mathematics, Korea Advanced Institute of Science and Technology, Daejeon, 305-701, South Korea.

ABSTRACT

Consider a model whose structure is representable via recursive graph or Bayes network and in which both observable and unobservable variables are involved. Assume that the variables are all binary and suppose that we want to predict about the unobservables based on the evidence from the observables. In the real world, we may not be able construct a model that exactly explains a certain phenomenon we are interested in. However, if we may give predictions in terms of class or category rather than probability, then we may not have to know the exact details of the model. It is shown in this article that when the observables are conditionally independent and conditionally stochastically ordered given the unobservables, the unobservables are conditionally ordered given the observables. This result suggests to some extent robustness of the conditional probabilities of the observables given the unobservables, and the simulation result strongly supports the suggestion.

Keywords: Agreement level, conditional probability, graphical model, basic structure, positive association.

1 Introduction and Motivation

The need for better understanding of knowledge states in the research fields of cognitive science and education calls for statistical technologies for linking behavioral outcomes to knowledge states. Some of the technologies are used in the form of graphical models ([8]) whose corresponding graphs include among others Bayesian network ([7, 3]), Markov network ([7]), and chain graph ([9, 17]).

Graphical models are useful in representing relationship among variables in terms of conditional independence. The corresponding graph is often called independence graph and edges are directed when the corresponding relationships can be interpreted as causal and not directed otherwise.

In educational testing, test results are used for predicting test takers' knowledge states, which are expressed in various forms such as total scores, estimates of person parameters or abstract ability parameters that are usually expressed by θ, probability of possessing a certain knowledge, and the likes ([12]). Ref. [18] considers a hybrid of graphical model and neural network ([6]) with a view to enhance the refinement level of knowledge representation.

When a graphical model is obtained with a corresponding independence graph, we can use it for prediction given a new observation. For instance, if we are interested in predicting test takers' knowledge states given their test results and the corresponding graphical model is obtained, then we can use it for predicting knowledge states of a test taker, assuming that all the variables are categorical, in terms of conditional probability given a test result of the test taker. A statistical technique useful for such prediction is what is called evidence propagation ([16]) and they are realized in computer programs such as HUGIN ([15]) and ERGO ([2]).

Consider a prediction problem where predictions are made in terms of categorical level rather than conditional probability. For instance, if we are interested in classifying a group of students into 5 subgroups with respect to states of a certain knowledge, we may use the conditional probability of possession of the knowledge and divide the student group in such a way that the top 20 % of the students belong to group 1, the next 20 % belong to group 2, and so on.

Since the classification is based on the rank of the conditional probabilities, the classification result is dependent upon the relative magnitude of the probability rather than the value of the probability itself. In this regard, when predictions are made for classification, the relative magnitude of probability, if available, may be of use as much as the actual magnitude of it.

This perspective leads us to the issue of robustness of the conditional probability to classification. Since we deal with relative magnitudes of the probability, the notion of stochastic ordering plays an important role in addressing the issue of robustness in classification. It is anticipated that the level of robustness varies according to the model structure. However, in order to see a possible range of robustness in prediction, a simulation experiment is carried out over a variety of models.

This paper is organized in 4 sections. Section 2 proposes a theorem that positive association among a set of binary variables preserves a stochastic ordering among the conditional probabilities of the binary variables. This result is carried over to section 3 in the form of simulated experiment in an effort to fathom the robustness of prediction which is made based on the ordering of the conditional probabilities of a given variable for a given data set. The simulation result shows a very high level of robustness when the variables are positively associated. Section 4 concludes the article with some further discussions.

2 Positive association and order preservation

We will write U for unobservable variables and X for observable variables. Vectors are bold-faced. 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 some $i = 1, \cdots, n$.

In a directed independence graph or a Bayes network such as the graph in Figure 1, if a pair of nodes are connected by an arrow, we call the node at the tail of the arrow a *parent* node of the node which is at the head of the arrow. For instance, in Figure 1, node U_1 is the only parent node of nodes U_2 and U_3, and node X_7 has three parent nodes U_4, U_5, and U_6. 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. If two nodes are linked by an arrow we say that the two nodes are neighbors.

Consider a pair of binary random variables U and X, taking on values 0 or 1, whose joint probability satisfies that

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

Then, we have

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

As a matter of fact, when $0 < P(U = 1) < 1$ and $0 < P(X = 1) < 1$, expression (2) is equivalent to

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

The left-hand side and the right-hand side of this inequality are, respectively, equal to

$$l = \frac{P(X = 1|U = 1)}{P(U = 1)P(X = 1|U = 1) + P(U = 0)P(X = 1|U = 0)}$$

and

$$r = \frac{P(X = 0|U = 1)}{P(U = 1)P(X = 0|U = 1) + P(U = 0)P(X = 0|U = 0)}.$$

Since $0 < P(U = 1) < 1$, we can see, by condition (1), that

$$l > 1 \text{ and } r < 1.$$

Thus (2) follows.

Under condition (1), we have

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

that is, U and X are positively associated. Actually, we can easily show that (1) and (2) are equivalent.

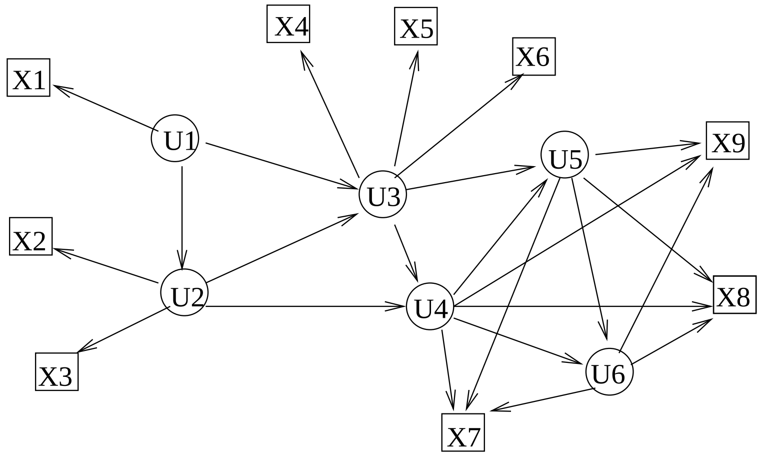


Figure 1: A Bayesian network where U variables are latent and X variables observable

We can extend this result to a situation where $X_i,\ i = 1, 2, \cdots, I$, are influenced by multiple U's, i.e., the conditional probability of X_i is subject to the states of some of $U_k,\ k = 1, 2, \cdots, K$ only. We may assume a recursive graph ([10]) or a directed independence graph ([8]) for $U_k,\ k = 1, 2, \cdots, K$ as in Figure 1. Note that since the conditional probability of X_i is subject to the states of $\{U_k\}_{k=1}^K$ only, $X_i,\ i = 1, 2, \cdots, I$, are independent given $U_k,\ k = 1, 2, \cdots, K$, as is illustrated in Figure 1.

Inequality (3) below is equivalent to that

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

This inequality says that $\mathbf{U} = \mathbf{v}$ is more likely than $\mathbf{U} = \mathbf{u}$ when $X_i = 1$ than when $X_i = 0$. But if we want to compare the likeliness of $U_k = 1$, the theorem below will help.

Theorem 1 *Let $\mathbf{X} = (X_1, \cdots, X_I)$ and $\mathbf{U} = (U_1, \cdots, U_K)$. Then the following two statements are equivalent.*
(i) For $i = 1, 2, \cdots, I$,

$$P(X_i = 1|\mathbf{u}) < P(X_i = 1|\mathbf{v}),\ \textit{when}\ \mathbf{u} \prec \mathbf{v}. \tag{3}$$

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

$$P(U_k = 1|\mathbf{x}) < P(U_k = 1|\mathbf{y}),\ \textit{when}\ \mathbf{x} \prec \mathbf{y}. \tag{4}$$

The strict inequality ($<$) in both (3) and (4) may be replaced by the plain inequality ($\leq$).

Proof: We remove the kth component of $\mathbf{U}$ and denote the resulting vector by $\mathbf{U}_{(k)}$. We will prove only that condition (i) implies (ii) since the proof for the other direction is by the same argument as for its counterpart except the switched role between $\mathbf{X}$ and $\mathbf{U}$.

$$\begin{aligned}
& P(U_k = 1|\mathbf{y}) - P(U_k = 1|\mathbf{x}) \\
= \ & \frac{P(U_k = 1)\sum_{\mathbf{u}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(\mathbf{y}|U_k = 1, \mathbf{u}_{(k)})}{\sum_{\mathbf{u}} P(\mathbf{u})P(\mathbf{y}|\mathbf{u})} \\
& - \frac{P(U_k = 1)\sum_{\mathbf{u}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(\mathbf{x}|U_k = 1, \mathbf{u}_{(k)})}{\sum_{\mathbf{u}} P(\mathbf{u})P(\mathbf{x}|\mathbf{u})} \\
= \ & \frac{b}{a} - \frac{d}{c},
\end{aligned} \tag{5}$$

where a, b, c, d are equal to the corresponding parts in expression (5).

$$\begin{aligned}
& \frac{bc - ad}{P(U_k = 1)} \\
& = \sum_{\mathbf{u}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(\mathbf{y}|U_k = 1, \mathbf{u}_{(k)}) \sum_{\mathbf{u}} P(\mathbf{u})P(\mathbf{x}|\mathbf{u}) \\
& \quad - \sum_{\mathbf{u}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(\mathbf{x}|U_k = 1, \mathbf{u}_{(k)}) \sum_{\mathbf{u}} P(\mathbf{u})P(\mathbf{y}|\mathbf{u}) \\
& = \sum_{\mathbf{u}_{(k)}} \sum_{\mathbf{v}} P(\mathbf{u}_{(k)}|U_k = 1)P(\mathbf{v}) \big(P(\mathbf{y}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{x}|\mathbf{v}) - P(\mathbf{x}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{y}|\mathbf{v})\big) \\
& = \sum_{\mathbf{u}_{(k)}} \sum_{\mathbf{v}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(V_k = 0, \mathbf{v}_{(k)}) \\
& \quad \times \big(P(\mathbf{y}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{x}|V_k = 0, \mathbf{v}_{(k)}) - P(\mathbf{x}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{y}|V_k = 0, \mathbf{v}_{(k)})\big) \\
& \quad + \sum_{\mathbf{u}_{(k)}} \sum_{\mathbf{v}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(V_k = 1, \mathbf{v}_{(k)}) \\
& \quad \times \big(P(\mathbf{y}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{x}|V_k = 1, \mathbf{v}_{(k)}) - P(\mathbf{x}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{y}|V_k = 1, \mathbf{v}_{(k)})\big)
\end{aligned} \tag{6}$$

Since $\mathbf{U}$ and $\mathbf{V}$ are the same in distribution, the second double summation in (6) is equal to zero, and we may rewrite the first double summation as

$$\begin{aligned}
& \sum_{\mathbf{u}_{(k)}} \sum_{\mathbf{v}_{(k)}} P(\mathbf{u}_{(k)}|U_k = 1)P(V_k = 0, \mathbf{v}_{(k)})P(\mathbf{y}|U_k = 1, \mathbf{u}_{(k)})P(\mathbf{x}|V_k = 0, \mathbf{v}_{(k)}) \\
- & \sum_{\mathbf{u}_{(k)}} \sum_{\mathbf{v}_{(k)}} P(\mathbf{v}_{(k)}|V_k = 1)P(U_k = 0, \mathbf{u}_{(k)})P(\mathbf{y}|U_k = 0, \mathbf{u}_{(k)})P(\mathbf{x}|V_k = 1, \mathbf{v}_{(k)}),
\end{aligned} \tag{7}$$

which takes on, when $\mathbf{x} \prec \mathbf{y}$, positive values by the assumption of the theorem.

If we replace the strict inequality in the condition of the theorem with $\leq$, then by applying the same argument as sbove, we can see from (7) that the strict inequality ($<$) be replaced in (4) with $\leq$. This completes the proof. □

It is worth noting that this theorem holds for any relationship among U_k as long as $0 < P(U_k = 1) < 1,\ k = 1, 2, \cdots, K$ and $0 < P(X_i = 1) < 1,\ i = 1, 2, \cdots, I$. When the $<$ and $\prec$ in expression (3) are replaced by $\leq$ and $\preceq$, respectively, the expression actually means that the conditional

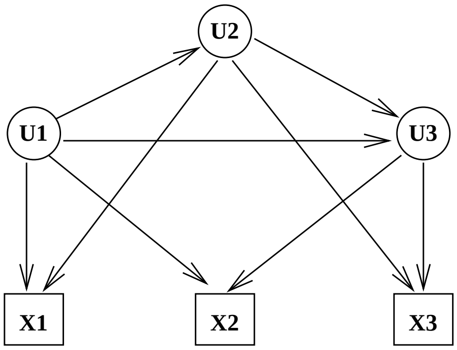


Figure 2: The recursive model considered in Example 2.1

distribution of X_i given $\mathbf{U} = \mathbf{u}$ is stochastically larger than that of X_i given $\mathbf{U} = \mathbf{v}$. Ref. [11] discusses properties concerning positive association among the X variables when the X variables are conditionally stochastically ordered given $\mathbf{U}$. Ref. [1] characterizes such X variables in more generic terms such as conditional association and vanishing conditional dependence.

Let $\mathbf{a}$ be a vector of 0's or 1's, and denote by $\delta(\mathbf{a})$ the number of 1-components in vector $\mathbf{a}$.

Theorem 2 *For $i, j = 1, 2, \cdots, I$, suppose that*

$$P(X_i = 1|\mathbf{u}) = P(X_j = 1|\mathbf{v}) \text{ whenever } \delta(\mathbf{u}) = \delta(\mathbf{v}).$$

Then we have, for $k = 1, 2, \cdots, K$, that

$$P(U_k = 1|\mathbf{x}) = P(U_k = 1|\mathbf{y}) \text{ whenever } \delta(\mathbf{x}) = \delta(\mathbf{y}).$$

Proof: We may apply the same argument as for the proof of Theorem 1 and show that the value of expression (7) equals zero. According to the condition of the theorem, X_i, $i = 1, 2 \cdots, I$, are symmetric in index. So in (7) we may exchange $\mathbf{x}$ and $\mathbf{y}$ with each other in the second double summation. From this follows the desired result. □

The condition of this theorem is very strong, but if the condition becomes milder as follows:

$$P(X_i = 1|\mathbf{u}) = P(X_i = 1|\mathbf{v}) \text{ and, for } i \neq j, P(X_i = 1|\mathbf{u}) \neq P(X_j = 1|\mathbf{v}) \tag{8}$$

when $\delta(\mathbf{u}) = \delta(\mathbf{v})$,

then the result of the theorem is not guaranteed as illustrated in the example below.

Example 2.1 Consider a recursive model of three binary latents, U_1, U_2, U_3, and three binary observables, X_1, X_2, X_3, whose relationship is depicted in Figure 2 and whose conditional probabilities are listed in Table 1. The probability model as in Table 1 satisfies condition (8).

Table 1: The conditional probabilities for the nodes in Figure 2

$P(U_1 = 1) = 0.8$	$P(X_1 = 1 \mid U_1 = 1, U_2 = 0) = 0.3$
	$P(X_1 = 1 \mid U_1 = 1, U_2 = 1) = 0.9$
$P(U_2 = 1 \mid U_1 = 0) = 0.15$	
$P(U_2 = 1 \mid U_1 = 1) = 0.85$	$P(X_2 = 1 \mid U_1 = 0, U_3 = 0) = 0.15$
	$P(X_2 = 1 \mid U_1 = 0, U_3 = 1) = 0.25$
$P(U_3 = 1 \mid U_1 = 0, U_2 = 0) = 0.1$	$P(X_2 = 1 \mid U_1 = 1, U_3 = 0) = 0.25$
$P(U_3 = 1 \mid U_1 = 0, U_2 = 1) = 0.15$	$P(X_2 = 1 \mid U_1 = 1, U_3 = 1) = 0.9$
$P(U_3 = 1 \mid U_1 = 1, U_2 = 0) = 0.15$	
$P(U_3 = 1 \mid U_1 = 1, U_2 = 1) = 0.9$	$P(X_3 = 1 \mid U_2 = 0, U_3 = 0) = 0.05$
	$P(X_3 = 1 \mid U_2 = 0, U_3 = 1) = 0.1$
$P(X_1 = 1 \mid U_1 = 0, U_2 = 0) = 0.1$	$P(X_3 = 1 \mid U_2 = 1, U_3 = 0) = 0.1$
$P(X_1 = 1 \mid U_1 = 0, U_2 = 1) = 0.3$	$P(X_3 = 1 \mid U_2 = 1, U_3 = 1) = 0.9$

Table 2: Values of $P(U_k = 1|\mathbf{x})$ as obtained from the model with the corresponding structure and the conditional probabilities given in Figure 2 and Table 1

	$P(U_k = 1\|\mathbf{x})$		
$\mathbf{x}$	$k = 1$	$k = 2$	$k = 3$
(1,0,0)	0.780	0.596	0.082
(0,1,0)	0.572	0.159	0.322
(0,0,1)	0.453	0.496	0.458
(1,1,0)	0.954	0.819	0.692
(1,0,1)	0.960	0.966	0.878
(0,1,1)	0.956	0.937	0.956

$P(U_k = 1|\mathbf{x})$ are listed in Table 2. For the configurations $\mathbf{x}$ with $\delta(\mathbf{x}) = 1$, we see different values of $P(U_k = 1|\mathbf{x})$, for each $k = 1, 2, 3$. Under condition (8), the values of $P(U_k = 1|\mathbf{x})$ are subject to the structural relationship between $\mathbf{U}$ and $\mathbf{X}$. Although the difference in value is subdued when $\delta(\mathbf{x}) = 2$, we also see a similar result for the configurations. □

Theorem 1 implies that, if we are interested in the ordering of $P(U_k = 1|\mathbf{x})$ for a set of $\mathbf{x}$ values that are totally ordered in terms of $\preceq$, we can order them without regard to the exact values of $P(X_i = 1|\mathbf{u})$. This fact throws light on the possibility that students may be ordered to some extent in accordance with the level of a certain ability without resort to the exact (mean) values of the probabilities of correct responses conditional on every possible state of the test-relevant abilities (denoted by $\mathbf{U}$). For instance, for each of the six abilities as symbolized by U_k in Figure 1, we can arrange a group of students in order of the ability level with some level of accuracy without knowing the exact values of $P(X_i = 1|\mathbf{u})$ where X_i symbolizes test-item scores. We can see in Table 2 that the inequality (4) holds when the $\mathbf{x}$ values are ordered.

In the section below we will see a simulation result on robustness of the class prediction when the classification is based on the conditional probability of the predicted variable.

3 A simulation result

Although we can not be fully optimistic on the robustness in prediction, Theorem 1 is a profound support for our optimism. To fathom the level of robustness, we will run a simulation experiment with a good deal of variations of the model in Figure 1. In the simulation, all the variables are binary, taking on 0 or 1, and we presuppose that we predict the states of U_k, $k = 1, \cdots, 6$, in terms of five levels of the $P(U_k = 1|\mathbf{X} = \mathbf{x})$, where the levels are based on the data for $\mathbf{X}$ to which a given model fits. For a set of data $\{\mathbf{x}_i\}_{i=1}^n$ with $\mathbf{x}_i = (x_{i1}, \cdots, x_{i9})$, we obtain a set of conditional probabilities $\{(P(U_k = 1|\mathbf{x}_i),\ k \in \{1, \cdots, 9\})\}_{i=1}^n$, and then arrange $\{P(U_k = 1|\mathbf{x}_i)\}_{i=1}^n$ in the descending order and rank them from the largest down. We then categorize the states of U_k $\{P(U_k = 1|\mathbf{x})\}_{k=1}^6$, for each k, into five levels from 1 through 5, 1 for the first 20% of the cases, 2 for the next 20% of the cases, and so on.

Suppose that U_k is the random indicator of the state of knowledge unit(KU) k which is good enough for a given task. Then we may interpret $P(U_k = 1)$ as the probability of possession of KU k that is required for the given task. So, from the view-point of subjective probability ([4]), we may regard the probability as a level of familiarity with the KU, 1 as a mastery level, 5 as a bottom level, and the intermediate values for the intermediate levels.

Our interest is in robustness of the prediction categories for the knowledge states. The probability model for the Bayesian network as in Figure 1 can be expressed as the product of $P(Z = z|pa(Z))$ where $pa(Z)$ denotes the set of the parent nodes of Z. $pa(Z)$ is empty when Z is a root node. The robustness is with respect to the set of the conditional probabilities $\{P(Z = z|pa(Z))\}_{Z\in\Psi}$ where Ψ is the set of all the variables contained in a given Bayesian network. We try many different values on the conditional probabilities under condition (3) and see how the prediction categories vary across a wide range of the values of the conditional probabilities.

For convenience' sake, we call by model 1 the model in Figure 1. For this model, we consider two versions of it. Version 1 is coated with a fixed set of numbers for the conditional probabilities $\{P(Z = z|pa(Z))\}_{Z\in\Psi}$ of the model while we use a wide range of numbers between 0.01 and 0.99 for the conditional probabilities in version 2. The values for $P(Z = 1|pa(Z)$ takes on 0 values only) have mean 0.12 with standard deviation 0.037, those for $P(Z = 1|pa(Z)$ are in perfect states) range from about 0.1 up to 0.99, and the values for the other imperfect states of $pa(Z)$ are selected so that condition (3) may be satisfied. The wide range for the perfect state of $pa(Z)$ is in an effort to reflect the real situation as much as possible. The fixed set of values for version 1 is listed in Table

Table 3: The (conditional) probabilities for the version 1 of model. In this table, we use V instead of U and X to represent a random variable. For convenience' sake, we write $P(V_3 = 1|v_1, v_2)$ for $P(V_3 = 1|V_1 = v_1, V_2 = v_2)$.

$P(V_1 = 1) = 0.55$
$P(V_2 = 1\|V_1 = 0) = 0.15$
$P(V_2 = 1\|V_1 = 1) = 0.65$
$P(V_3 = 1\|v_1, v_2) = 0.1, \quad v_1 + v_2 = 0$
$P(V_3 = 1\|v_1, v_2) = 0.15, \quad v_1 + v_2 = 1$
$P(V_3 = 1\|v_1, v_2) = 0.65, \quad v_1 + v_2 = 2$
$P(V_4 = 1\|v_1, v_2, v_3) = 0.1, \quad v_1 + v_2 + v_3 = 0$
$P(V_4 = 1\|v_1, v_2, v_3) = 0.15, \quad v_1 + v_2 + v_3 = 1$
$P(V_4 = 1\|v_1, v_2, v_3) = 0.25, \quad v_1 + v_2 + v_3 = 2$
$P(V_4 = 1\|v_1, v_2, v_3) = 0.65, \quad v_1 + v_2 + v_3 = 3$
$P(V_5 = 1\|v_1, v_2, v_3, v_4) = 0.1, \quad v_1 + v_2 + v_3 + v_4 = 0$
$P(V_5 = 1\|v_1, v_2, v_3, v_4) = 0.15, \quad v_1 + v_2 + v_3 + v_4 = 1$
$P(V_5 = 1\|v_1, v_2, v_3, v_4) = 0.25, \quad v_1 + v_2 + v_3 + v_4 = 2$
$P(V_5 = 1\|v_1, v_2, v_3, v_4) = 0.35, \quad v_1 + v_2 + v_3 + v_4 = 3$
$P(V_5 = 1\|v_1, v_2, v_3, v_4) = 0.65, \quad v_1 + v_2 + v_3 + v_4 = 4$
$P(V_6 = 1\|v_1, v_2, v_3, v_4, v_5) = 0.1, \quad v_1 + v_2 + v_3 + v_4 + v_5 = 0$
$P(V_6 = 1\|v_1, v_2, v_3, v_4, v_5) = 0.15, \quad v_1 + v_2 + v_3 + v_4 + v_5 = 1$
$P(V_6 = 1\|v_1, v_2, v_3, v_4, v_5) = 0.25, \quad v_1 + v_2 + v_3 + v_4 + v_5 = 2$
$P(V_6 = 1\|v_1, v_2, v_3, v_4, v_5) = 0.35, \quad v_1 + v_2 + v_3 + v_4 + v_5 = 3$
$P(V_6 = 1\|v_1, v_2, v_3, v_4, v_5) = 0.45, \quad v_1 + v_2 + v_3 + v_4 + v_5 = 4$
$P(V_6 = 1\|v_1, v_2, v_3, v_4, v_5) = 0.65, \quad v_1 + v_2 + v_3 + v_4 + v_5 = 5$

3. Of course, as for version 1 also, the values for $\{P(Z = z|pa(Z))\}_{Z \in \Psi}$ must satisfy condition (3).

In the simulation, we compare the class predictions between the two versions of a model. Suppose that the predictions are made for N students who took a given test for which model 1 is appropriate with 9 test items and 6 KUs that are relevant to the test. For the given model, we denote by Y_{ijk} the predicted class by version i for student j regarding the level of KU k and let $D_{jk} = Y_{2jk} - Y_{1jk}$. Since the classes are labelled 1 through 5, we have that $-4 \leq D_{jk} \leq 4$. From the distribution of D, we can obtain a measure of agreement between the two sets of predictions.

For knowledge unit k, we can obtain the relative frequencies of D through the expression

$$r_{kd} = \frac{\text{the number of cases that } D_{jk} = d}{N}.$$

These r_{kd} values are obtained for each set of conditional probabilities for version 2 of a model. And by generating 50 different sets of conditional probabilities for version 2, the average of the 50

Table 4: $\bar{r}_k$ values for model 1

U_i	$\bar{r}_{-4}$	$\bar{r}_{-3}$	$\bar{r}_{-2}$	$\bar{r}_{-1}$	$\bar{r}_0$	$\bar{r}_1$	$\bar{r}_2$	$\bar{r}_3$	$\bar{r}_4$	$\sum_{d=-1}^{1} \bar{r}_d$
U_1	0.000	0.000	0.003	0.106	0.734	0.142	0.015	0.000	0.000	0.982
U_2	0.000	0.000	0.002	0.109	0.727	0.153	0.009	0.000	0.000	0.989
U_3	0.000	0.000	0.000	0.066	0.809	0.123	0.001	0.000	0.000	0.998
U_4	0.000	0.000	0.005	0.096	0.743	0.148	0.008	0.000	0.000	0.987
U_5	0.000	0.005	0.035	0.082	0.702	0.143	0.033	0.000	0.000	0.927
U_6	0.000	0.001	0.005	0.074	0.795	0.115	0.009	0.000	0.000	0.984
average	0.000	0.001	0.008	0.089	0.752	0.137	0.012	0.000	0.000	0.978

accruing r_{kd} values was obtained for each k (denote it by $\bar{r}_k$). The averages for model 1 are listed in Table 4.

It is indicated in Table 4 that on average the predictions from the two versions exactly agree for about 75% of the cases (see the column of $\bar{r}_0$ in the table) and the last column of the table shows that almost all the predictions agree up to one class level. The average of the 6 values in the last column is 0.978. This suggests that almost all of the predictions for the knowledge levels may be different only up to one class level.

Further simulations were carried out to investigate how the agreement level varies across the model structures. It is usually the case that KUs are related each other according to their intrinsic nature and that multiple KUs are required for a test item to be solved. For example, in Figure 1, it is indicated that the state of KU 3 is influenced by the state of KUs 1 and 2 and similarly for the other KUs and that test item 1 requires KU 1 only while test item 7 requires KUs 4, 5, and 6. In this figure, 6 items tap only one KU and the other three tap three KUs.

When the KUs are linked together in a model, the conditional probability of a KU depends upon all the item score variables (X's). For example, although U_1 in Figure 1 is linked directly to X_1 only, $P(U_1 = 1|x_1, \cdots, x_9) \neq P(U_1 = 1|x_1)$. It looks likely that X_1 influences most upon the conditional probability $P(U_1 = 1|x_1, \cdots, x_9)$, but it is time-consuming to compute the amount of influence. The influence upon the conditional probability of a node is subject to the model structure. If U_1 is isolated from the other U variables and is tapped by X_1 only, then $P(U_1 = 1|x_1)$ takes on two different values since X_1 is binary. Thus, under condition (3), the class predictions must agree between the two versions for any values of $P(X_1 = 1|u_1)$. Such a perfect agreement is not guaranteed in general when U is tapped by multiple X's.

We will investigate into the agreement level not over model structures but over an individual U variable with a variety of situations of neighborhood nodes some of which are U variables and others are X variables. Some basic situations of neighborhoods are listed in Figure 3. Since our predictions

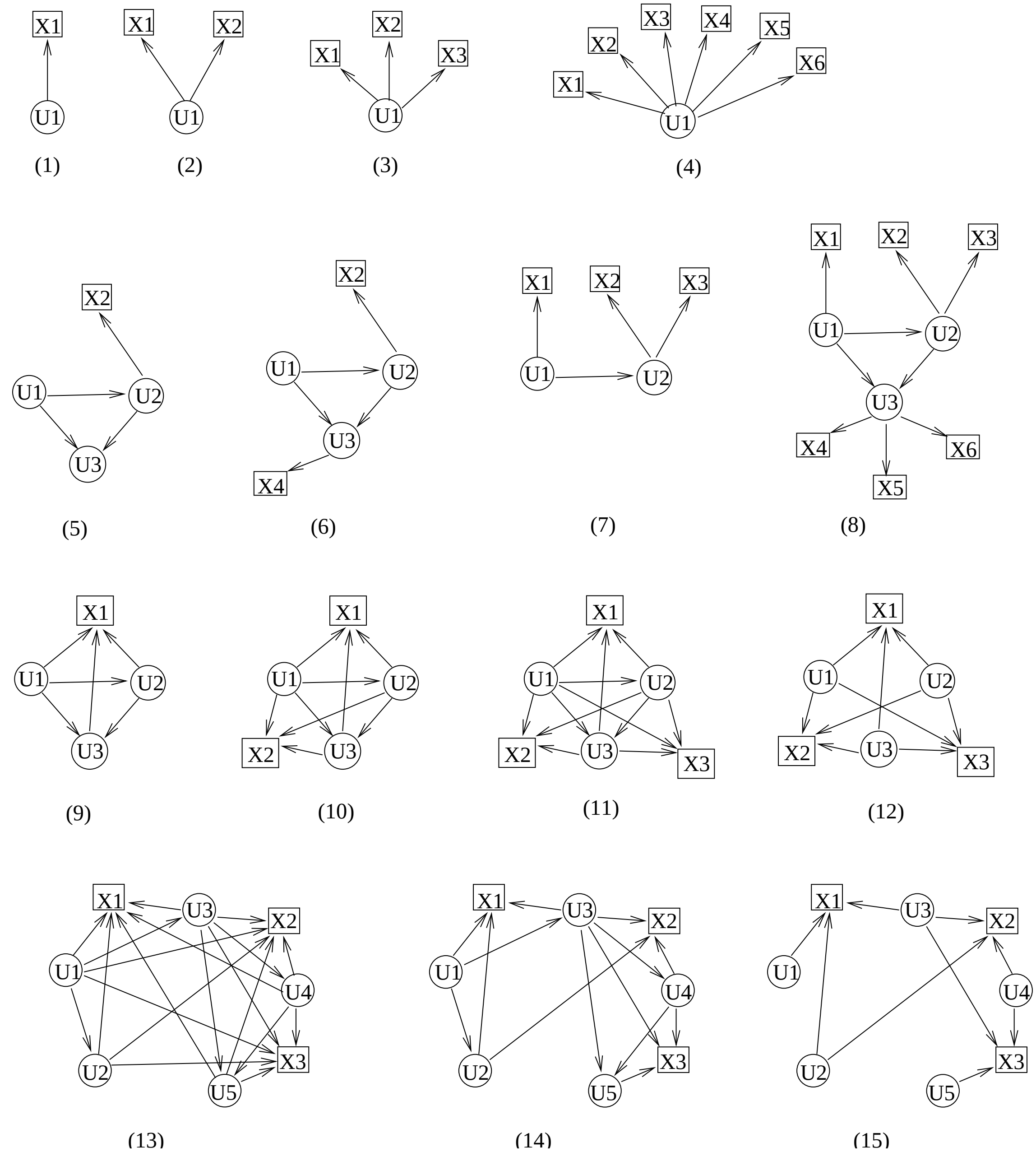


Figure 3: Basic model structures that are considered for the simulation study

are based on $P(U_k = 1|\mathbf{x})$, it is important to look into how X variables are linked to a given U variable. In panels 1 through 4, U_1 is a single parent node of one or multiple X variables, while, in panels 5 and 6, U_1 is not a parent of any X variables but is neighbored with other U variables. We will call the type of neighborhood situation as in the first four panels $N1$-situation (here "N" stands for "neighborhood") and the situation as in panels 5 and 6 $N2$-situation. Panels 7 and 8 are taken into consideration to see how the prediction robustness is affected when U_1 is a parent of X variables and also neighbored with some U variables. We will call this neighborhood situation $N3$-situation. In a sense, $N3$-situation is a mixture of $N1$ and $N2$-situations. This type of tapping is commonplace in educational testing, but we pay attention to these basic situations to explore how the neighborhood-situations affect the prediction robustness.

In these three neighborhood-situations, each X variable has only one parent node as a U variable. But, in panels 9 through 15, all the X variables have multiple parents. In panels 9 through 12, 14 and 15, every X variable has three parent nodes, while every X variable has five parent nodes in panel 13. When U is one of multiple parents of an X variable, we will say that the U is in an $N4$-situation. A common feature in panels 9 through 13 is that all the U variables are equally tapped by X variables, but it is not the case for panels 14 and 15. In panel 14, U_1 is linked directly to X_1 only and linked to X_2 and X_3 through other U variables.

We define the agreement level up to class-difference j in predictions for variable U_k by

$$\alpha_j(U_k) = \sum_{|d| \leq j} r_{kd}.$$

In particular, $\alpha_0(U_k)$ is the exact agreement level for U_k. When confusion is not likely, we will simply use α_j instead of $\alpha_j(U_k)$.

It is displayed in Table 5 that $\alpha_1(U_1)$ values are larger than or equal to 0.97 when U_1 is in $N1$, $N2$, or $N3$-situation. And it is also the case in the $N4$ situations when the U variable is linked to other U variables. In panels 12 and 15, all the U variables are marginally independent each other, and it is interesting to see that the α_1 values for U_1, U_2, and U_3 in panel 12 are all high around 0.95 while those for $U_1, \cdots, U_5$ in panel 15 are 0.86, 0.97, 0.98, 0.96, and 0.85, respectively. In panel 15, U_1 and U_5 each are linked directly to only one X variable, and their α_1 values are around 0.85, while the levels for the other U variables in panels 12 and 15 are all larger than or equal to 0.95.

In panel 15, U_1 shares its only child X_1 with two other U variables, U_2 and U_3, which have two or three child nodes each, and all the U variables are marginally independent. The marginal independence implies that the U variables are free to take on any values from their support sets and so the influence of U_1 upon X_1 is in general less than the influence when there is no marginal independence among the U variables. If U_1, U_2, and U_3 are associated among themselves so highly

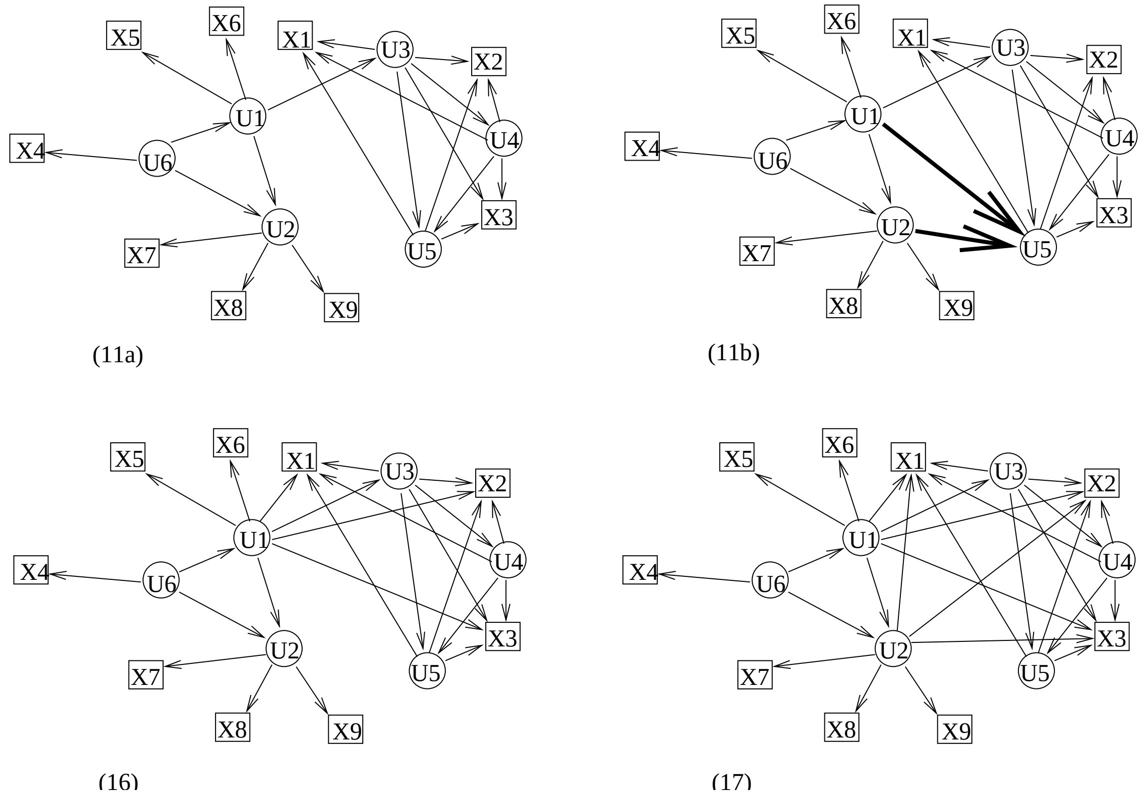
Table 5: α_1 and α_0 values for the neighborhood situations as listed in Figures 3 and 4

Panel	U_1		U_2		U_3		U_4		U_5		U_6	
	α_1	α_0	α_1	α_0	α_1	α_0	α_1	α_0	α_1	α_0	α_1	α_0
1	1.00	1.00										
2	1.00	0.91										
3	0.98	0.84										
4	0.97	0.71										
5	1.00	1.00	1.00	1.00	1.00	1.00						
6	1.00	0.90	1.00	1.00	1.00	1.00						
7	1.00	0.95	1.00	0.95								
8	0.99	0.79	0.99	0.80	1.00	0.81						
9	1.00	1.00	1.00	1.00	1.00	1.00						
10	1.00	0.89	1.00	0.89	1.00	0.89						
11	0.99	0.80	0.99	0.78	1.00	0.80						
12	0.95	0.69	0.95	0.69	0.96	0.72						
13	0.98	0.83	0.97	0.83	0.98	0.83	0.98	0.84	0.98	0.83		
14	1.00	0.93	0.97	0.84	0.99	0.83	0.98	0.88	1.00	0.98		
15	0.86	0.58	0.97	0.68	0.98	0.70	0.96	0.66	0.85	0.56		
	$(U_2^8)^*$		(U_3^8)		(U_1^{11})		(U_2^{11})		(U_3^{11})		(U_1^8)	
11a	0.99	0.73	1.00	0.81	0.99	0.74	0.93	0.70	0.98	0.80	0.98	0.73
11b	0.98	0.71	0.99	0.70	0.99	0.77	1.00	0.79	0.99	0.75	0.99	0.74
16	0.99	0.74	0.93	0.57	0.97	0.69	0.94	0.55	0.95	0.53	0.98	0.73
17	0.96	0.68	0.91	0.58	0.94	0.55	0.85	0.38	0.85	0.39	0.93	0.67

*: Panels (11a) and (11b) are for comparison with the basic structures in panels (8) and (11). U_i^j stands for the node U_i in panel (j). The α values in the last four rows of the table are comparable with the α values for the nodes as indicated in the fifth row from the bottom of the table.

that we may regard them as one variable, α_1 and α_0 values may be as large as those for U_1 in panel 11. We can see this phenomenon by comparing the α values between panels 11 and 12 and between panels 14 and 15. Note that the difference is more obvious in the α_0 values. The simulation result strongly indicates that high association among a set of U variables may yield larger α values for the individual U variable in the set than for the U variables which are less associated among themselves.

Figure 4 displays four graphs. Panel (11a) is a combined graph of panels (8) and (11) with only one edge added between U_1 and U_3. Thus it makes sense to compare the α values between panel (11a) and its counter part. In the table, the row of U_i^j's which are explained at a footnote to the table is inserted for the convenience of readership. For instance, in panel (11a), variables U_1 and U_3 correspond respectively to U_2 in panel (8) and U_1 in panel (11). The α_1 values are almost the same between the variables of each corresponding pair, while α_0 values are slightly smaller for panel (11a). Panel (11b) was considered to see how additional edges affect the α values, and no special trend is seen between panels (11a) and (11b). The values are more or less the same between the two



(11a)

(11b)

(16)

(17)

Figure 4: Some more model structures as an extension to the model structures in Figure 3. The thick arrows in panel (11b) are the only addition to the graph in panel (11a).

panels.

Panels (16) and (17) are obtained by further adding edges to the graph in panel (11a). Note that variables $X_1, \cdots, X_3$ each has four and five parent nodes respectively in panels (16) and (17). We can see a trend in the α values that they decrease as we move across panels in the order of (11a), (16), and (17). The α_0 values decreased by considerable amounts while the α_1 values did not change much. Fortunately, for most of the cases in Table 5, the α_1 values were close to 1 and none of the values were smaller than 0.85. In other words, the classification were very robust up to class level 1.

In a nutshell, the simulation result strongly shows a general trend that, as is implicitly manifested in the panels (1) through (4) and in the pair of panels (13) and (14), as a U variable becomes a parent of more X variables, the α values for it gets smaller. Another important phenomenon is that, as is observed among others in the pair of panels (11) and (12) and the pair of panels (14) and (15), as the U variables become more associated each other, the α values tend to increase. These two phenomena are apparent in every model structure that we have considered in this section.

4 Concluding remarks

When we are in a situation of predicting for some variable in terms of class level, it is desirable that we look into the model structure regarding conditional independence and conditional stochastic ordering. Theorem 1 says that when predictions are to be made on U variables and the conditional independence and conditional stochastic ordering hold for X variables given $\mathbf{U}$, the predictions are robust to some extent if they are given in terms of class level rather than probability.

We have considered some "basic" structures for investigating the prediction robustness. The size of the set of all the possible model structures increases exponentially with the number of variables involved in the model. So it is almost impossible to consider all of them to find the range of prediction robustness, nor is easy to find a measure of the robustness for a particular model structure. As a matter of fact, such a task seems not necessary at all.

Through a simulation experiment that is focused on an individual U variable and its neighborhood in the model structure, we have seen a good level of robustness expressed in terms of α_1 values. A crucial point of view is that the influence of a set of X variables upon the prediction for a U variable may be determined by how the X variables are connected to the U variable and how the U variable is connected to other U variables.

In the experiment, we considered the model structures where a U variable is a parent node of up to five X variables. It is possible that the α_1 value falls below 0.85 for a U variable. If such is the case, we may exclude the U variable from our robust prediction scheme, or we may give predictions for the variable with the corresponding α_1 value. We are now at a position to recommend a robust prediction approach as follows:

> Given a Bayesian network of binary variables, suppose that we are to predict for every U variable in the model in terms of class-level. Provided that there is no variable in the model that has more than five parent variables, we may use the probability values as in Table 3 or some values near by them. Then apply the method as in the simulation experiment in the previous section to obtain the α_0 and α_1 values for all the U variables in the model. Finally, give predictions for the subjects in the real data of the Bayesian network.

There is no easy way of finding such a table as Table 3 when a given model is complicated involving a large number of variables which may often be the case in the real world. Actually Table 3 was found after a sequence of trial and error experiments and it worked well for a variety of model structures.

Although we have considered Bayesian networks only, the prediction robustness would be valid for other forms of graphical models such as Markov networks ([7]) and mixture of the two types. One of the reasons is that these three forms of graphs all share the Markov properties that are essentially transformable into the same graphical layout ([13], [14]).

The result of this article is relevant to all the problems where predictions may be made based on the relative magnitudes of the conditional probabilities under the assumption that the variables are binary and are positively associated each other. The robust prediction or decision making scheme will save much of our time and effort provided that the assumptions are valid.

References

[1] B.W. Junker and J.L. Ellis, A characterization of monotone unidimensional latent variable models, The Annals of Statistics, **25**, 3, 1327-1343 (1997).

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

[3] F.V. Jensen, An Introduction to Bayesian Networks (Springer-Verlag, NewYork, 1996).

[4] H.E. Jr. Kyburg and H.E. Smokler, Studies in Subjective Probability (Edited) (Robert E. Krieger Publishing Company, Huntington, New York, 1980).

[5] J.A. Anderson, An Introduction to Neural Networks (The MIT Press, 1995).

[6] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of The National Academy of Sciences, **79**, 2554-2558 (1982).

[7] J. Pearl, Probabilistic Reasoning In Intelligent Systems: Networks of Plausible Inference (Morgan Kaufmann, San Mateo, CA, 1988).

[8] J. Whittaker, Graphical Models in Applied Multivariate Statistics (Wiley, New York, 1990).

[9] M. Frydenberg, The chain graph Markov property, Scandinavian Journal of Statistics, **17**, 333-353 (1990).

[10] N. Wermuth and S.L. Lauritzen, Graphical and recursive models for contingency tables, Biometrika, **70**, 3, 537-552 (1983).

[11] P.W. Holland and P.R. Rosenbaum, Conditional association and unidimensionality in monotone latent variable models, The Annals of Statistics, **14**, 4, 1523-1543 (1986).

[12] R.J. Mislevy, Evidence and inference in educational assessment, Psychometrika, **59**, 4, 439-483 (1994).

[13] S.A. Andersson, D. Madigan and M.D. Perlman, On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs, Scandinavian Journal of Statistics, **24**, 81-102 (1997a).

[14] S.A. Andersson, D. Madigan and M.D. Perlman, A characterization of Markov equivalence classes for acyclic digraphs. The Annals of Statistics, **25**, 2, 505-541 (1997b).

[15] S.K. Andersen, F.V. Jensen, K.G. Olesen and F. Jensen, HUGIN: A shell for building Bayesian belief universes for expert systems [computer program] (HUGIN Expert Ltd., Aalborg, Denmark,1989).

[16] S.L. Lauritzen and D.J. Spiegelhalter, Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems, J. of Royal Statistical Soc. B, **50**, 2, 157-224 (1988).

[17] S.L. Lauritzen and N. Wermuth, Graphical models for associations between variables, some of which are qualitative and some quantitative, The Annals of Statistics, **17**, 1, 31-57 (1989).

[18] S.P. Marshall, Schemas in Problem Solving (Cambridge University Press, 1995).