Mixed Non-Parametric Continuous and Discrete Bayesian Belief Nets

Anca Hanea
Delft Institute for Applied Mathematics
Delft University of Technology
Mekelweg 4, 2628 CD Delft
The Netherlands
A.Hanea@ewi.tudelft.nl

Dorota Kurowicka
Delft Institute for Applied Mathematics
Delft University of Technology
Mekelweg 4, 2628 CD Delft
The Netherlands
D.Kurowicka@ewi.tudelft.nl

Abstract

This paper introduces mixed non-parametric continuous and discrete Bayesian Belief Nets (BBNs) using the copula-vine modelling approach. We extend the theory for non-parametric continuous BBNs to include ordinal discrete random variables; that is variables which can be written as monotone transforms of uniforms. The dependence structure among the variables is given in terms of (conditional) rank correlations. We use an adjusted rank correlation coefficient for discrete variables, and we emphasise the relationship between the rank correlation of two discrete variables and the rank correlation of their underlying uniforms. The approach presented in this paper is illustrated by means of an example.

1 Introduction

Applications in various domains often lead to high dimensional dependence modelling. Decision makers and problem owners are becoming increasingly sophisticated in reasoning with uncertainty. This motivates the development of generic tools, which can deal with two problems that occur throughout applied mathematics and engineering: uncertainty and complexity.

Graphical models provide a general methodology for approaching these problems. A bayesian belief net is one of the probabilistic graphical models, which encodes the probability density or mass function of a set of variables by specifying a number of conditional independence statements in a form of an acyclic directed graph and a set of probability functions. The visual representation can be very useful in clarifying previously opaque assumptions about the dependencies between different variables. Our focus is on mixed non-parametric continuous and discrete BBNs.

In a non-parametric continuous BBN, nodes are associated with arbitrary continuous invertible distribution functions and arcs with (conditional) rank correlations, which are realised by a copula with the zero independence property (Kurowicka and Cooke 2004). The (conditional) rank correlations assigned to the edges are algebraically independent, and there are tested protocols for their use in structured expert judgement (Morales et al. 2007). We note that quantifying BBNs in this way also requires assessing all (continuous, invertible) one dimensional marginal distributions. On the other hand, the dependence structure is meaningful for any such quantification, and need not be revised if the univariate distributions are changed.

We extend this approach to include ordinal discrete random variables which can be written as monotone transforms of uniform variates, perhaps taking finitely many values. The dependence structure, however, must be defined with respect to the uniforms. The rank correlation of two discrete variables and the rank correlation of their underlying uniforms are not equal. Therefore one needs to study the relationship between these two rank correlations.

The paper is organised as follows: Section 2 introduces the details of non-parametric continuous BBNs using the normal copula vine modelling approach presented in Hanea et al. (2006). In order to extend this approach to include ordinal discrete random variables, an adjusted rank correlation coefficient for such variables is defined. Section 3 presents a correction for the population version of Spearman's rank

correlation coefficient r for discrete random variables, and describes the relationship between the rank correlation of two discrete variables and the rank correlation of their underlying uniforms (Hanea et al. 2007). For a better understanding of the methodology described here, an application model is presented in Section 4. Finally, Section 5 presents conclusions and recommendations for future work.

2 Non-Parametric Continuous BBNs

A continuous non-parametric BBN is a directed acyclic graph, together with a set of (conditional) rank correlations and a set of marginal distributions. Nodes are associated with arbitrary continuous invertible distribution functions and arcs with constant (conditional) rank correlations that are realised by a copula for which (conditional) correlation 0 entails (conditional) independence[1] (Kurowicka and Cooke 2004). For each variable i with parents $i_1 ... i_{p(i)}$, we associate the arc $i_{p(i)-k} \longrightarrow i$ with the conditional rank correlation:

$$\begin{cases} r(i,i_{p(i)}), & k=0 \\ r(i,i_{p(i)-k}|i_{p(i)},\ldots,i_{p(i)-k+1}), & 1 \leq k \leq p(i)-1. \end{cases}$$

The assignment is vacuous if $\{i_1 ... i_{p(i)}\} = \emptyset$ (see Figure 1).

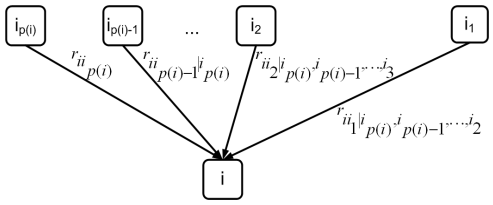


Figure 1: Node i of a BBN and the set of parent nodes for i.

Therefore, every arc in the BBN is assigned a (conditional) rank correlation between parent and child. These assignments are algebraically independent. Moreover they uniquely determine the joint distribution. The proof of this fact is available in Kurowicka and Cooke (2006) and it is based on the close relationship between non-parametric BBNs and another graphical model, namely vines (Cooke 1997; Bedford and Cooke 2002).

A way of stipulating a joint distribution is by sampling it, hence we use a sampling protocol based on vines to specify and analyse the BBN structure. As we already mentioned, the (conditional) rank correlations assigned to the arcs of a BBN can be realised by a copula with the zero independence property. Unfortunately, for sampling a large BBN structure with a general copula, extra calculations may be required. These calculations consist of numerical evaluations of multiple integrals, which are very time consuming. This disadvantage vanishes when using the normal copula (Hanea et al. 2006). Hence we will present the sampling algorithm for non-parametric continuous BBNs with the normal copula. Let us start with a BBN on n variables $X_1, \ldots X_n$, with continuous, invertible distribution functions $F_1, \ldots, F_n$. We transform $X_1, \ldots X_n$ to the standard normal variables $Y_1, \ldots Y_n$ via the transformation $Y_i = \Phi^{-1}(F_i(X_i))$, $(\forall i)(i = 1, \ldots, n)$, where Φ is the cumulative distribution function of the standard normal distribution. Since $\Phi^{-1}(F_i(X_i))$ are strictly increasing transformations, the same (conditional) rank correlations correspond to the pairs of transformed variables $Y_1, \ldots Y_n$. Further, since all conditional distributions are normal we can use Pearson's transformation (Pearson 1907) to calculate $\rho_{i,j|D} = 2\sin(\frac{\pi}{6} \cdot r_{i,j|D})$, where $r_{i,j|D}$ is the conditional rank correlation between Y_i and one of its parents, Y_j, given a subset of parents, denoted by D (see Figure 1). For normal variables, conditional and partial correlations are equal.

[1] Such copulae are said to have the *zero independence property.*

The relationship between vines and continuous non-parametric BBNs ensures that there is a unique joint normal distribution for $Y_1, \dots Y_n$ satisfying the partial correlation specifications. Moreover there is a unique corespondent correlation matrix R (Bedford and Cooke 2002). One can calculate the correlation matrix R, using the recursive formula for partial correlations (Yule and Kendall 1965):

$$\rho_{12;3,\dots,n} = \frac{\rho_{12;4,\dots,n} - \rho_{13;4,\dots,n} \cdot \rho_{23;4,\dots,n}}{((1-\rho^2_{13;4,\dots,n}) \cdot (1-\rho^2_{23;4,\dots,n}))^{\frac{1}{2}}}. \tag{2.1}$$

We can now sample the joint normal distribution of $Y_1, \dots, Y_n$, with correlation matrix R (Tong 1990) and for each sample $((y_1^j), (y_2^j), \dots, (y_n^j))$, $j = 1, ..., N$, calculate: $(F_1^{-1}(\Phi(y_1^j)), F_2^{-1}(\Phi(y_2^j)), \dots, F_n^{-1}(\Phi(y_n^j)))$.

In this way we realise the joint distribution of the initial variables $X_1, \dots X_n$, together with the specified dependence structure.

We intend to use the same protocol in the case of mixed non-parametric continuous and discrete BBNs. Hence, we will further consider BBNs whose nodes represent both discrete and continuous variables. We are interested in discrete ordinal variables which can be written as monotone transforms of uniform variables. This should be understood as follows: let X be a discrete variable with m possible values $x_1, ..., x_m$ each with probability $p_1, ..., p_m$, respectively. We call U_X the underlying uniform for the discrete variable X, if $P(U_X < \sum_{j=1}^{k} p_j) = \sum_{j=1}^{k} p_j$, $k = 1, ..., m$. The dependence structure in the BBN must be defined with respect to the underlying uniform variables. The rank correlation of 2 discrete variables and the rank correlation of their underlying uniforms are not equal, hence one needs to establish the relationship between them.

3 Spearman's Rank Correlation for Ordinal Discrete Random Variables

Before defining the rank correlation of 2 discrete variables, we will first recall the definition of the population version of Spearman's rank correlation coefficient, in terms of the probabilities of concordance and discordance (e.g., Nelsen 1999).

Consider a population distributed according to 2 variates X and Y. Two members (X_1, Y_1) and (X_2, Y_2) of the population will be called *concordant* if:

$$X_1 < X_2, Y_1 < Y_2 \text{ or } X_1 > X_2, Y_1 > Y_2.$$

They will be called *discordant* if:

$$X_1 < X_2, Y_1 > Y_2 \text{ or } X_1 > X_2, Y_1 < Y_2.$$

The probabilities of concordance and discordance are denoted with P_c, and P_d respectively. The population version of Spearman's r is defined as proportional to the difference between the probability of concordance, and the probability of discordance for two vectors (X_1, Y_1) and (X_2, Y_2), where (X_1, Y_1) has distribution F_{XY} with marginal distribution functions F_X and F_Y and X_2, Y_2 are independent with distributions F_X and F_Y; moreover (X_1, Y_1) and (X_2, Y_2) are independent (e.g., Joe 1997):

$$r = 3 \cdot (P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]). \tag{3.1}$$

The above definition is valid only for populations for which the probabilities of $X_1 = X_2$ and $Y_1 = Y_2$ are zero. The main types of such populations are an infinite population with both X and Y distributed continuously, or a finite population where X and Y have disjoint ranges (Hoffding 1947).

In order to formulate the population version of Spearman's rank correlation coefficient r, for discrete random variables, one needs to correct for the probabilities of $X_1 = X_2$ and $Y_1 = Y_2$. This correction is derived in Hanea et al. (2007). In this section we present the main results.

Let us consider the discrete random vectors (X_1, Y_1), (X_2, Y_2), where X_2 and Y_2 are independent with the same marginal distributions as X_1 and Y_1, respectively; moreover (X_1, Y_1) and (X_2, Y_2) are independent. The states of X_i are ranked from 1 to m; the states of Y_i are ranked from 1 to n.[2] The joint probabilities of (X_1, Y_1) and (X_2, Y_2) are given in terms of p_{ij} and q_{ij}, $i = 1, .., m; j = 1, .., n$, respectively.

Table 1: Joint distribution of (X_1, Y_1) (left); Joint distribution of (X_2, Y_2) (right)

$X_1 \setminus Y_1$	1	2	...	n	
1	p_{11}	p_{12}	...	p_{1n}	p_{1+}
2	p_{21}	p_{22}	...	p_{2n}	p_{2+}
...	...	...	...	...	...
m	p_{m1}	p_{m2}	...	p_{mn}	p_{m+}
	p_{+1}	p_{+2}	...	p_{+n}	

$X_2 \setminus Y_2$	1	2	...	n	
1	q_{11}	q_{12}	...	q_{1n}	p_{1+}
2	q_{21}	q_{22}	...	q_{2n}	p_{2+}
...	...	...	...	...	...
m	q_{m1}	q_{m2}	...	q_{mn}	p_{m+}
	p_{+1}	p_{+2}	...	p_{+n}	

where p_{i+}, $i = 1, ..., m$ represent the margins of X_1 and X_2; and the margins of Y_1 and Y_2 are denoted p_{+j}, $j = 1...n$. One can rewrite each q_{ij} as $q_{ij} = p_{i+}p_{+j}$, for all $i = 1, ..., m$, and $j = 1, ..., n$. Using this terminology we calculate the difference between the probabilities of concordance and discordance as follows:

$$P_c - P_d = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}\left(\sum_{k\neq i}\sum_{l\neq j} sign(k-i)(l-j)q_{kl}\right)\right) \tag{3.2}$$

The adjusted rank correlation coefficient of two discrete variables is given by the following theorem:

Theorem 3.1. *Consider a population distributed according to two variates X and Y. Two members (X_1, Y_1) and (X_2, Y_2) of the population are distributed as in Table 1. Let $P_c - P_d$ be given by formula (3.2). Then the population version of Spearman's rank correlation coefficient of X and Y is:*

$$\bar{r} = \frac{P_c - P_d}{\sqrt{\left(\sum_{j>i} p_{i+}p_{j+} - \sum_{k>j>i} p_{i+}p_{j+}p_{k+}\right)\cdot\left(\sum_{j>i} p_{+i}p_{+j} - \sum_{k>j>i} p_{+i}p_{+j}p_{+k}\right)}}$$

As we already mentioned, the discrete distributions which interest us are the ones that can be obtained as monotone transforms of uniform variables. These distributions can be constructed by only specifying the marginal distributions and a copula. Each term p_{ij} from Table 1 (left) can be written in terms of the chosen copula, as follows[3]:

$$p_{ij} = C\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) + C\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right) - C\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - C\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right) \tag{3.3}$$

Each copula can be parameterised by its rank correlation r, so we will use the notation C_r instead of C.

The main result for our approach is given by the following theorem, namely the relationship between the rank correlation of the discrete variables and the rank correlation of the uniform variates.

[2] In Section 2 we used the letter n for the number of variables in a BBN. We now use the same letter, in a different context, without any connection with the previous use.

[3] It is worth mentioning that by using this construction we do not obtain all possible joint distributions, given the margins.

Theorem 3.2. *Let C_r be a copula and (X,Y) a random vector distributed as in Table1(left), where each p_{ij} is given by formula (3.3) . Then the rank correlation of X and Y is denoted $\bar{r}_C$ and it has the same expression as $\bar{r}$, where:*

$$P_c - P_d = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\left(p_{i+} + p_{(i+1)+}\right)\left(p_{+j} + p_{+(j+1)}\right) C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} p_{i+}p_{+j} \tag{3.4}$$

Moreover, if C_r is a positively ordered copula (Nelsen 1999), then $\bar{r}_C$ is an increasing function of the rank correlation of the underlying uniforms.

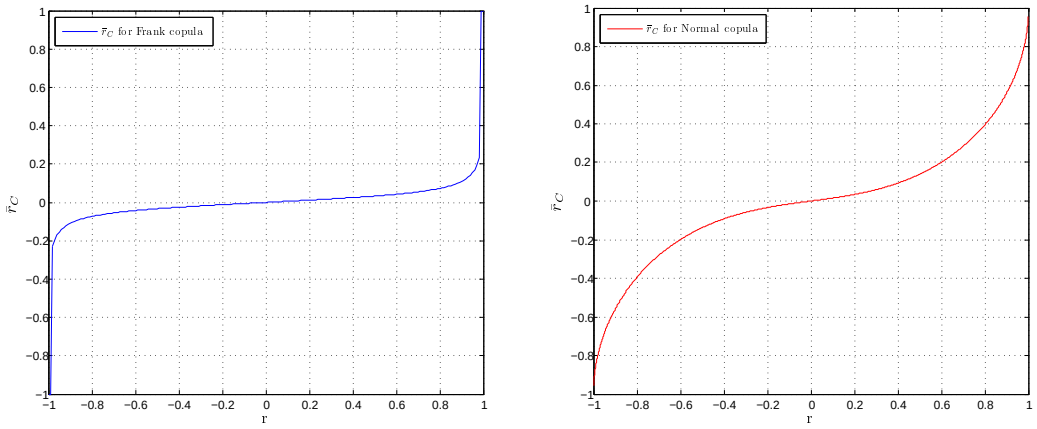
Note that any copula can be used in expression 3.4 from Theorem 3.2. If the independent copula is used, the equation simplifies to zero, as expected.

In contrast with the continuous case, the adjusted coefficient for the discrete variables is a function of not only the copula, but also the marginal distributions.

Examples. *We will further investigate the relationship between $\bar{r}_C$ and the dependence parameter, r, of the copula. We choose different copulae (with more emphasis on the normal copula) and different marginal distributions for 2 discrete random variables X and Y.*

If we consider 2 ordinal responses X and Y, both uniformly distributed across a small number of states, $\bar{r}_C$ and r tend to be very similar, for any choice of a positive ordered copula. Moreover $\bar{r}_C$ covers the whole range of r. Increasing the number of states for X and Y, makes $\bar{r}_C$ approximately equal[4] to r.

When marginal distributions are not uniform, the relationship changes. Figure 2 presents the relationship between r and $\bar{r}_C$, for 2 discrete variables X and Y, with 3 states each. Their marginal distributions are the same, namely[5]: $p_{1+} = p_{+1} = 0.01$; $p_{2+} = p_{+2} = 0.98$ and $p_{3+} = p_{+3} = 0.01$. We use Frank's copula to obtain Figure 2a, and the normal copula in Figure 2b.



(a) $p_{1+} = p_{+1} = p_{3+} = p_{+3} = 0.01, p_{2+} = p_{+2} = 0.98$. The joint distribution is constructed using Frank's copula.

(b) $p_{1+} = p_{+1} = p_{3+} = p_{+3} = 0.01, p_{2+} = p_{+2} = 0.98$. The joint distribution is constructed using the Normal copula.

Figure 2: The relationship between the parameter r, of a chosen copula, and the adjusted rank correlation $\bar{r}_C$, for discrete random variables with equal and symmetric marginal distributions.

As both Frank's copula and the normal copula are positively ordered, $\bar{r}_C$ is an increasing function of r. Since the marginal distributions are symmetric, the range of rank correlations realised for the discrete

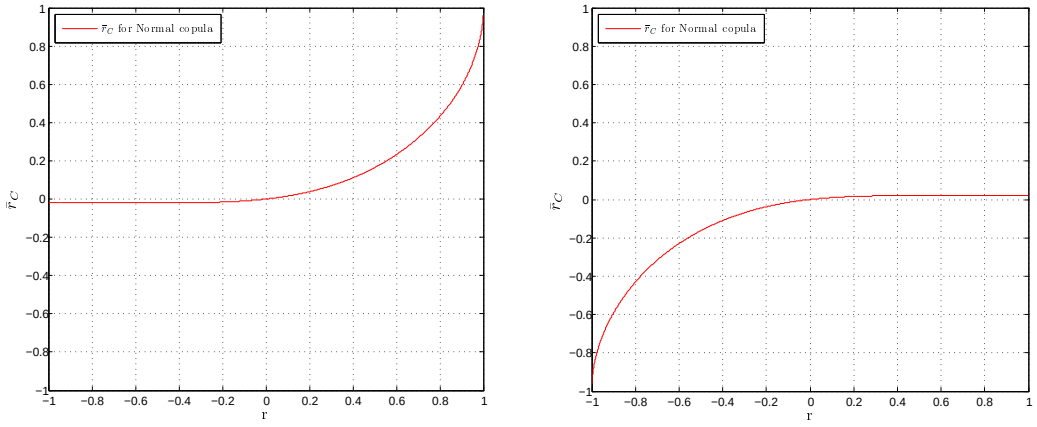
[4] 10 states for each variable will suffice to obtain differences of order 10^{-3}, between $\bar{r}_C$ and r.

[5] We use the same notation as in Table 1 to describe the marginal distributions of X and Y.

variables is the entire interval $[-1, 1]$. Notice that the relationship is very nonlinear. This strong nonlinearity is caused by the choice of $p_{2+} = p_{+2} = 0.98$.

If we now consider variables with identical, but not symmetric marginal distributions, the relationship is not symmetric around 0 anymore. In this case the whole range of positive dependence can be attained, but the range of negative association is bounded below, as shown in Figure 3a.

We will further consider marginal distributions that are not identical, but "complementary", in the sense that: $p_{1+} = p_{+3}$; $p_{2+} = p_{+2}$ and $p_{3+} = p_{+1}$. Then the entire range of negative association is possible, but the range of positive association is bounded above, as shown in Figure 3b.



(a) $p_{1+} = p_{+1} = p_{2+} = p_{+2} = 0.01, p_{3+} = p_{+3} = 0.98$.

(b) $p_{2+} = p_{3+} = p_{+1} = p_{+2} = 0.01, p_{1+} = p_{+3} = 0.98$.

Figure 3: The relationship between the parameter r, of the Normal copula, and the adjusted rank correlation $\bar{r}_C$, for discrete random variables with equal (a), and "complementary" (b) marginal distributions.

Further, if variables X and Y have 3 states, such that $p_{1+} = 0.01$, $p_{2+} = 0.98$, $p_{3+} = 0.01$ (for X) and $p_{+1} = 0.19$, $p_{+2} = 0.01$, $p_{+3} = 0.80$ (for Y), we can observe (see Figure 4a) that both positive and negative dependencies are bounded.

One can also calculate bounds for $\bar{r}_C$, by using the Frechet bounds for C_r in expression 3.4. These bounds are shown in Figure 4a. Since we know the bounds, we can normalise the rank coefficient $\bar{r}_C$, such that it covers the entire interval $[-1, 1]$. The result of this normalisation is displayed in Figure 4b.

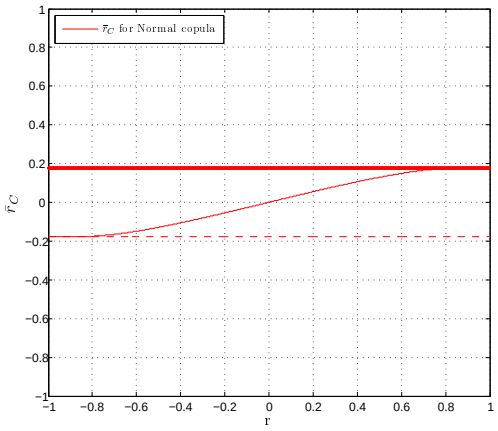
These results provide a better understanding of the techniques involved in modelling dependence between discrete random variables, using copula functions.

4 Illustrations

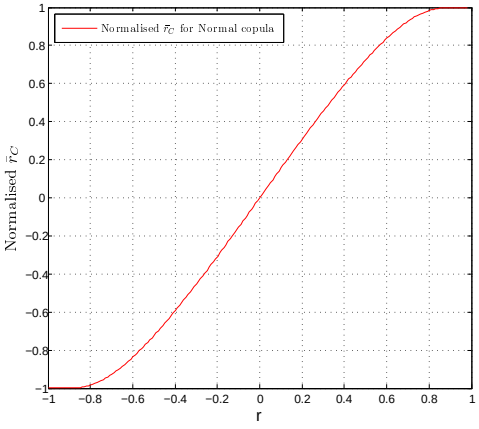
We will further illustrate the concepts and methods described until now, with an example. This example is loosely based on a project undertaken by the European Union. The name of the project is Beneris (which stands for Benefit and Risk) and it focuses on the analysis of health benefits and risks associated with food consumption[6]. The model introduced here is a highly simplified version of the BBN model used in the project (Jesionek and Cooke 2007). The goal is to estimate the beneficial and harmful health effects in a specified population, as a result of exposure to various contaminants and nutrients through ingestion of fish. Figure 5a resembles the version of the model that we are considering for purely illustrative purposes.

The variables of interest for this model are the health endpoints resulting from exposure to fish constituents, namely cancer and cardiovascular risk. These risks are defined in terms of remaining lifetime

[6]http://www.beneris.eu/

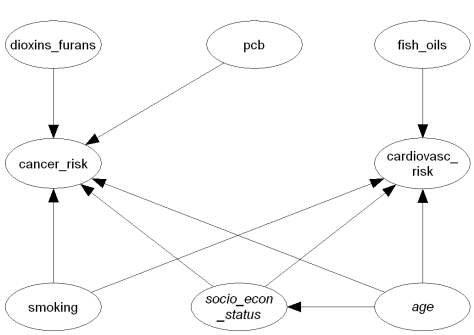


(a) The relation between r and $\bar{r}_C$, for X and Y, with not uniform, not equal, not "complementary" marginal distributions.

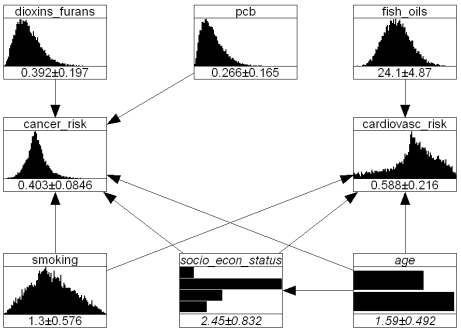


(b) The relation between r and the normalised $\bar{r}_C$, for X and Y, with not uniform, not equal, not "complementary" marginal distributions.

Figure 4: The relation between the parameter r, of the Normal copula, and $\bar{r}_C$ (a); the relation between r of the Normal copula, and the normalised adjusted rank correlation $\bar{r}_C$ (b).



(a) Simplified fish consumption BBN.



(b) Simplified fish consumption BBN with histograms.

Figure 5: Simplified Bayesian Belief Net for fish consumption risks.

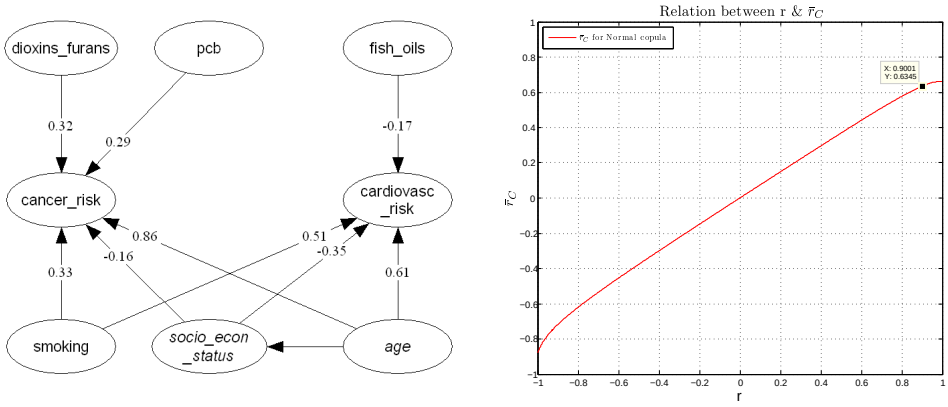
risks.

The 3 fish constituents that are considered are: dioxins/furans, polychlorinated biphenyls, and fish oil. The first two are persistent and bio-accumulative toxins which cause cancer in humans. Fish are a significant source of exposure to these chemicals. Fish oil is derived from the tissues of oily fish and has high levels of omega-3 fatty acids which regulate cholesterol and reduce inflammation throughout the human body.

Moreover, personal factors such as smoking, socioeconomic status and age may influence cancer and cardiovascular risk. Smoking is measured as yearly intake of nicotine during smoking and passive smoking, while the socioeconomic status is measured by income, which is represented by a discrete variable with 4 income classes. The age is taken, in this simplified model, as a discrete variable with 2 states, 15 to 34 years, and 35 to 59 (we are considering only a segment of the whole population).

The distributions of the variables that form the BBN are presented in Figure 5b. They are chosen by the authors for illustrative purposes only. As we already mentioned there are 2 discrete (age and socioeconomic status), and 6 continuous random variables. Some indication of the relationships between

variables is given in their description above. For example, the personal factors: smoking and age will be positively correlated with both risks, whereas the socioeconomic status will be negatively correlated with cancer and cardiovascular risk. The (conditional) rank correlations assigned to the arcs of the BBN must be gathered from existing data or expert judgement (Morales et al. 2007). In this example, the numbers are, again, chosen by the authors. Figure 6a presents the same BBN, only now (conditional) rank correlations are assigned to each arc, except one.



(a) Simplified fish consumption BBN with (conditional) rank correlations.

(b) The relation between the parameter r, of the Normal copula, and $\bar{r}_C$.

Figure 6: Simplified Bayesian Belief Net for fish consumption risks; (conditional) rank correlations are assigned to the arcs of the BBN.

The arc between the 2 discrete variables *"age"* and *"soci_econ_status"* is not assigned any rank correlation coefficient. Let us assume that the correlation between them can be calculated from data, and its value is 0.63. As we stressed in the previous sections, the dependence structure in the BBN must be defined with respect to the underlying uniform variables. Hence, we first have to calculate the rank correlation of the underlying uniforms, r, which corresponds to $\bar{r}_C = 0.63$. In doing so, we use the normal copula. The relationship between r and $\bar{r}_C$ is shown in Figure 6b. Therefore, one must assign the rank correlation 0.9 to the arc of the BBN, in order to realise a correlation of 0.63 between the discrete variables. To double check this, we can sample the structure, using the protocol described in Section 2, and calculate the sample rank correlation matrix (see Table2).

Table 2: The sample rank correlation matrix.

	dioxins furans	pcb	fish oils	smoking	socioecon. status	age	cancer risk	cardiovasc. risk
dioxins/furans	1	-0.0002	-0.0021	0.0012	0.0013	0.0012	0.322	0.0014
pcb	-0.0002	1	0.0033	0.0008	-0.0015	-0.0011	0.2718	-0.001
fish oils	-0.0021	0.0033	1	0.0015	-0.0007	-0.0022	-0.0006	-0.1654
smoking	0.0012	0.0008	0.0015	1	0.0018	0.0005	0.2953	0.501
socioecon. status	0.0013	-0.0015	-0.0007	0.0018	1	***0.6376***	-0.124	-0.2684
age	0.0012	-0.0011	-0.0022	0.0005	***0.6376***	1	0.1348	-0.0554
cancer risk	0.322	0.2718	-0.0006	0.2953	-0.124	0.1348	1	0.5391
cardiovasc. risk	0.0014	-0.001	-0.1654	0.501	-0.2684	-0.0554	0.5391	1

Similarly, we can choose the required correlations between a uniform variable underlying a discrete, and other continuous variables (e.g. the uniform underlying *"age"*, and *"cardiaovasc_risk"*). The theory for this is in development, but not yet rigourously proven. For this example, we simply discretised the continuous variable in a large number of states, and proceed as for 2 discrete random variables.

Figures 5 and 6a are obtained with a software application, called Unicorn[7]. Unicorn allows for quantification of mixed non-parametric continuous and discrete BBNs (Kurowicka and Cooke 2006; Ababei et al. 2007). Once the BBN is specified, via the marginal distributions and the (conditional) rank correlations, the structure can be sampled. Moreover, evidence can be propagated through the graph, via analytical conditioning (Hanea et al. 2006). One or more of the variables can be set to a point value within their range, and the BBN can then be updated.

Let us return to the fish consumption example. We will further examine the situation in which there is a very high risk of cancer. We will conditionalise on the 0.9 value of cancer risk. Figure 7 presents how this information affects the other variables in the graph.

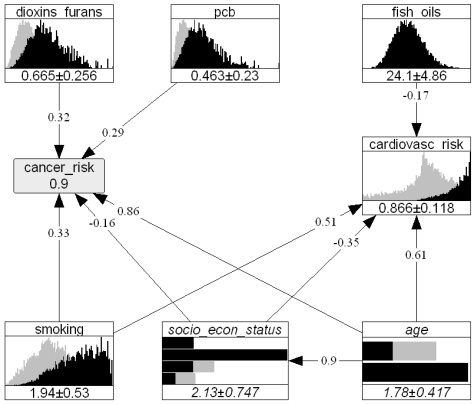


Figure 7: Conditionalised BBN for *cancer_risk = 0.9*.

The grey distributions in the background are the unconditional marginal distributions, provided for comparison. The conditional means and standard deviations are displayed under the histograms. Figure 7 summarises the combination of factors that increases the risk of cancer to 0.9. From the shift of the distributions, one can notice that if a person is neither very young, nor very wealthy, smokes a lot, and ingests more dioxins/furans, and polychlorinated biphenyls, is more likely get cancer. Because some of this factors influence also the cardiovascular risk, the shift in their distributions causes an increase in the cardiovascular risk as well.

5 Conclusions

We have shown how mixed non-parametric continuous and discrete BBNs can be modelled with the copula-vine approach. We have extended the theory for continuous BBNs to include discrete random variables that can be written as monotone transforms of uniform variables. In this approach, the dependence structure must be defined - via (conditional) rank correlations - with respect to the uniform variates. We have described the relationship between the rank correlation of two discrete variables ($\bar{r}_C$) and the rank correlation of their underlying uniforms (r).

Once $\bar{r}_C$ for 2 discrete variables is obtained, we use this relationship to calculate the rank correlation of their respective underlying uniforms. A value for $\bar{r}_C$ can be either obtained from data, or form experts. The technique for eliciting (conditional) rank correlations for discrete variables is still an open issue.

A mixed non-parametric continuous and discrete BBN will contain arcs that connect discrete nodes with continuous nodes. Hence, correlations between uniforms underlying the discrete variables, and other continuous variables, will be also required. A rigourous proof for the theoretical results in this direction is under development.

[7]A light version of Unicorn is available at http://dutiosc.twi.tudelft.nl/ risk/index.php.

References

Ababei, D. A., D. Kurowicka, and R. Cooke (2007). Uncertainty analysis with unicorn. In *Proceedings of the Third Brazilian Conference on Statistical Modelling in Insurance and Finance*.

Bedford, T. J. and R. Cooke (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics 30*(4), 1031–1068.

Cooke, R. (1997). Markov and entropy properties of tree and vine-dependent variables. In *Proceedings of the Section on Bayesian Statistical Science, American Statistical Association*.

Hanea, A. M., D. Kurowicka, and R. Cooke (2006). Hybrid method for quantifying and analyzing bayesian belief nets. *Quality and Reliability Engineering International 22*(6), 613–729.

Hanea, A. M., D. Kurowicka, and R. Cooke (2007). The population version of spearman's rank correlation coefficient in the case of ordinal discrete random variables. In *Proceedings of the Third Brazilian Conference on Statistical Modelling in Insurance and Finance*.

Hoffding, W. (1947). On the distribution of the rank correlation coefficient r when the variates are not independent. *Biometrika 34*, 183–196.

Jesionek, P. and R. Cooke (2007). Generalized method for modeling dose-response relations application to beneris project. Technical report, European Union project.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Kurowicka, D. and R. Cooke (2004). Distribution - free continuous bayesian belief nets. Proceedings Mathematical Methods in Reliability Conference.

Kurowicka, D. and R. Cooke (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley.

Morales, O., D. Kurowicka, and A. Roelen (2007). Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering and System Safety, Article in Press, Accepted Manuscript*.

Nelsen, R. (1999). *An Introduction to Copulas*. Lecture Notes inStatistics. New York: Springer- Verlag.

Pearson, K. (1907). Mathematical contributions to the theory of evolution. *Biometric Series. VI.Series*.

Tong, Y. (1990). *The Multivariate Normal Distribution*. New York: Springer-Verlag.

Yule, G. and M. Kendall (1965). *An introduction to the theory of statistics*. Belmont, California: Charles Griffin & Co. 14th edition.