INVITATION

Algorithms for Non-Parametric Bayesian Belief Nets

You are warmly welcome to attend the defence of my Ph.D. thesis and propositions on Monday 15 December 2008 at 10:00am in the Senaatszaal of the Auditorium of the Delft University of Technology, Mekelweg 5, Delft.

Prior the defence, at 09:30am, there will be a short presentation for non-experts.

The defence will be followed by a reception in the Auditorium.

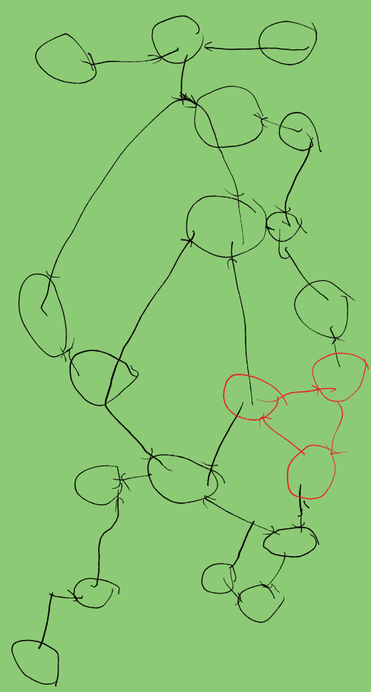
Hierbij nodig ik u uit voor de openbare verdediging van mijn proefschrift op maandag 15 december 2008 om 10:00 uur in de Senaatszaal in de Aula van de Technische Universiteit Delft, Mekelweg 5, Delft.

Voorafgaand aan de verdediging geef ik om 09:30 uur een korte toelichting op mijn onderzoek.

Na afloop van de verdediging bent u van harte welkom bij de receptie op dezelfde locatie.

Anca Hanea

ALGORITHMS
FOR
NON-PARAMETRIC
BAYESIAN BELIEF NETS



Anca Hanea

Algorithms for Non-Parametric Bayesian Belief Nets

Anca Hanea

Propositions

accompanying the thesis

Algorithms for Non - Parametric Bayesian Belief Nets

Anca Hanea

1. A non-parametric continuous BBN is a way of factorising the determinant of the correlation matrix and also a way of decomposing the mutual information.

2. The population version of Spearman's rank correlation for the case of ordinal discrete random variables proposed in Chapter 3 of this thesis coincides with the one derived by Neslehova (2007). In the particular case of binary variables, the alternative form of Spearman's rank correlation proposed by Vandenhende et al. (2003), and the normalized correction for the population version of Spearman's rank correlation proposed here are identical.

   J.Neslehova, On Rank Correlation Measures for Non-Continuous Random Variables, Journal of Multivariate Analysis, 98, 3, 544-567, 2007
   F. Vandenhende, P. Lambert , Improved Rank-based Dependence Measures for Categorical Data , Statistics and Probability Letters, 63, 157-163, 2003

3. Causal information about the data can be represented better in a non-parametric continuous BBN than in a simple regression model. This is particularly true in situations where the set of regressors have individually weak correlations with the predicted variable, but they are collectively important.

4. Non-parametric continuous BBNs typically exhibit conditional variances that are not constant, contrary to what standard regression models assume.

5. Vines provide a flexible way to model multivariate data with complex patterns of dependence in the tails, and are often superior in this regard to other models for capturing high dimensional dependence.

   D.Berg,K.Aas, Models for construction of multivariate dependence: A comparison study, Forthcoming in The Europen Journal of Finance, 2008
   M. Fischer , C. Kck, S. Schlter, F. Weigert, Multivariate Copula Models at Work: Outperforming the "desert island copula"? Discussion Paper 2007.

6. Mixed discrete & non-parametric continuous BBNs can handle hundreds of variables (Morales et al., 2007). Expert judgement is often essential in quantifying such models. If experts are treated as statistical hypotheses this need not damage the objectivity of the BBN model.

7. Non-parametric continuous BBNs with other than the normal copula may be employed in cases where the graphical structure does not contain large undirected cycles.

8. Supporting literature on parameter assessment in classical Gaussian BBN models is difficult to find. Direct communication with the members of the community UAI has shed precious little light on the matter.

9. People assume that time is a strict progression of cause to effect...but actually, from a non-linear, non-subjective viewpoint, it is more like a big ball of wibbly-wobbly...timey-wimey...stuff.

   The Doctor

10. While most of us can see only a few have the gift of sight.

    The Cat Empire

These propositions are considered opposable and defendable and as such have been approved by the supervisor, Prof. Dr. R.M.Cooke.

Stellingen

behorende bij het proefschrift

Algorithms for Non - Parametric Bayesian Belief Nets

Anca Hanea

1. Een niet-parametrische continue BBN is een manier om de determinant van een correlatiematrix te factoriseren en geeft tevens een decompositie van de mutual information.

2. De in Hoofdstuk 3 voorgestelde populatieversie van de Spearman rangcorrelatie voor ordinale discrete stochasten komt overeen met die door Neslehova (2007) is afgeleid. In het speciale geval van binaire variabelen vallen de alternatieve vorm van de Spearman correlatie voorgesteld door Vandenhende et al. (2003) en de genormaliseerde correctie voor de populatie versie die in dit proefschrift is voorgesteld, samen.

   J.Neslehova, On Rank Correlation Measures for Non-Continuous Random Variables, Journal of Multivariate Analysis, 98, 3, 544-567, 2007
   F. Vandenhende, P. Lambert , Improved Rank-based Dependence Measures for Categorical Data , Statistics and Probability Letters, 63, 157-163, 2003

3. Causale informatie met betrekking tot de gegevens kan beter voorgesteld worden in een niet-parametrische BBN dan in een eenvoudig regressiemodel. Dit geldt met name in gevallen waarin de regressoren een sterke onderlinge correlatie vertonen, terwijl de correlatie met de afhankelijke variabele zwak is.

4. Niet-parametrische continue BBNs vertonen doorgaans een niet-constant voorwaardelijke variantie in tegenstelling tot de gangbare veronderstellingen bij regressiemodellen.

5. Vines verschaffen een flexibele manier om multivariate gegevens met complexe patronen van startafhankelijkheid te modelleren, en zijn vaak superior in dit opzicht aan andere modellen voor hoog-dimensionale afhankelijkheid.

   D.Berg,K.Aas, Models for construction of multivariate dependence: A comparison study, Forthcoming in The Europen Journal of Finance, 2008
   M. Fischer , C. Kck, S. Schlter, F. Weigert, Multivariate Copula Models at Work: Outperforming the "desert island copula"? Discussion Paper 2007.

6. Gemengde discrete en niet-parametrisch continue BBNs kunnen honderden variabelen aan (Morales et al. 2007). Expert-mening is menigmaal essentieel bij het quantificeren van zulke modellen. Wanneer experts als statistische hypothesen worden behandeld, hoeft dit de objectiviteit van de BBN niet te schaden.

7. Niet-parametrisch continue BBNs met andere copulae dan de normale kunnen gebruikt worden, als de grafische structuur geen grote gerichte cycli heeft.

8. Ondersteunende literatuur voor het schatten van parameters in de klassieke Gaussische BBNs is zeer moeilijk vindbaar. Directe communicatie met leden van de betreffende onderzoeksgemeenschap UAI heeft opvallend weinig aan het licht gebracht.

9. Mensen veronderstellen dat de tijd een strikte progressie is van oorzaak naar gevolg...maar vanuit een niet-linear, niet-subjectief gezichtspunt, lijkt tijd meer op een grote bal van wibbly-wobbly...timey-wimey dingen.

   The Doctor

10. Terwijl de meesten van ons kunnen zien, hebben slechts weinig de gave van het zien.

    The Cat Empire

Deze stellingen worden opponeerbaar en verdedigbaar geacht en zijn als zodanig goedgekeurd door de promotor, Prof. Dr. R.M.Cooke.

ALGORITHMS FOR NON - PARAMETRIC BAYESIAN BELIEF NETS

ALGORITHMS FOR NON - PARAMETRIC BAYESIAN BELIEF NETS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 15 december 2008 om
10:00 uur

door

Anca Maria HANEA

Master of Science in Applied Mathematics
geboren te Bucureşti, România.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. R.M. Cooke

Copromotor: Dr. D. Kurowicka

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. R.M. Cooke	Technische Universiteit Delft, promotor
Dr. D. Kurowicka	Technische Universiteit Delft, copromotor
Prof. dr. C. Czado	Technische Universität München
Prof. dr. H. Joe	University of British Columbia, Vancouver
Prof. dr. C. Genest	Université Laval, Québec
Prof. dr. L.J.M. Rothkrantz	Netherlands Defense Academy
Prof. dr. ir. G. Jongbloed	Technische Universiteit Delft
Prof. dr. F.M. Dekking	Technische Universiteit Delft, reservelid

ISBN:
Typeset by the author with the LaTeX Documentation System.
On the cover: drawing by Alexandru Hanea
Printed in The Netherlands by: Wöhrmann Print Service

Pentru Mama

Contents

Chapter 1

Introduction

High dimensional probabilistic modelling using graph theory is employed in several scientific fields, including statistics, physics, biology and engineering. Graphical models proved to be a flexible probabilistic framework, and their use has increased substantially, hence the theory behind them has been constantly developed and extended. They merge graph theory and probability theory to provide a general setting for models in which a number of variables interact. The graphical structure is a collection of vertices (nodes) and links. The visual representation can be very useful in clarifying previously opaque assumptions about the dependencies between different variables. Each node in the graph represents a random variable. The links represent the qualitative dependencies between variables. The absence of a link between two nodes means that any dependence between these two variables is mediated via some other variables. Graphical models are used for probabilistic inference, decision making and data mining, in large-scale models in which a multitude of random variables are linked in complex ways.

There are two main types of graphical models: directed and undirected. The directed ones are based on directed acyclic graphs and their use can be tracked back to the pioneering work of Wright (1921). The graphical models with undirected links are generally called Markov random fields or Markov networks. Further we shall use the term *edge* for an undirected link, and *arc* for a directed link. Hybrid models are also available; they include both arcs and edges (Lauritzen 1996). Directed graphs and undirected graphs make different statements of conditional independence, therefore there are probability distributions that are captured by a directed graph and are not captured by any undirected graph, and conversely (Pearl 1988).

We restrict our attention to the directed graphical models called Bayesian belief nets, also known as belief networks, Bayesian networks, probabilistic networks, causal networks, and knowledge maps. We shall use the name Bayesian belief net and the abbreviation BBN. Among the reasons for choosing BBNs to represent high dimensional distributions we mention their capability of displaying relationships among variables in an intuitive manner, and that of representing cause-effect

relationships through the directionality of the arcs. Moreover, in contrast with Markov networks, they can represent induced and non-transitive dependencies[1]. A very important feature of a BBN is that it can be used for inference. One can calculate the distributions of unobserved nodes, given the values of the observed ones. If the reasoning is done "bottom-up" (in terms of the directionality of arcs), the BBN is used for diagnosis, whereas if it is done "top-down", the BBN serves for prediction.

1.1 Bayesian Belief Nets - Facts and Fiction

Bayesian Belief Nets (BBNs) are directed acyclic graphs. The nodes of the graph represent univariate random variables, which can be discrete or continuous, and the arcs represent direct influences[2].

BBNs provide a compact representation of high dimensional uncertainty distributions over a set of variables $(X_1, ..., X_n)$ (Cowell et al. 1999; Pearl 1988) and encode the probability density or mass function on $(X_1, ..., X_n)$ by specifying a set of conditional independence statements in a form of an acyclic directed graph and a set of probability functions.

From basic probability theory we know that every joint density, or mass function can be written as a product:

$$f(x_1, x_2, ..., x_n) = f(x_1) \prod_{i=2}^{n} f(x_i | x_1 ... x_{i-1}) . \tag{1.1.1}$$

Note that specifying this joint mass or density involves specifying values of an n-dimensional function. The directed graph of a BBN induces a (generally non-unique) ordering, and stipulates that each variable is conditionally independent of all predecessors in the ordering given its direct predecessors. The direct predecessors of a node i, corresponding to variable X_i are called *parents* and the set of all i's parents is denoted $Pa(i)$. Figure 1.1 shows a very simple BBN on 4 variables: X_1, X_2, X_3, and X_4, where X_1, X_2, X_3 form the set $Pa(4)$; X_4 is called a *child* of X_1, X_2, X_3.

Each variable is associated with a conditional probability function of that variable given its parents in the graph, $f(X_i | X_{Pa(i)}), i = 1, \ldots, n$. The conditional independence statements encoded in the graph allow us to simplify the expression of the joint probability from (1.1.1) as follows:

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f\left(x_i | x_{Pa(i)}\right) . \tag{1.1.2}$$

[1] A node with converging arrows is a configuration that yields independence in Markov networks and dependence in BBNs.

[2] BBNs can also contain functional nodes, i.e nodes which are functions of other nodes. The ensuing discussion refers to probabilistic nodes.

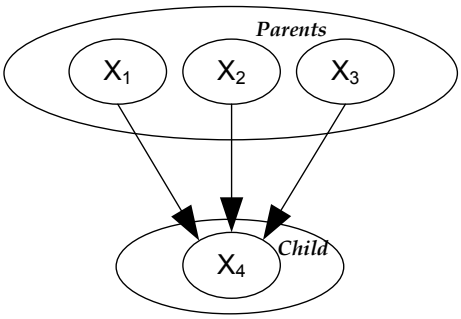


Figure 1.1: *A BBN on 4 variables.*

If $Pa(i) = \emptyset$, node i is called a *source node* and $f\left(x_i|x_{Pa(i)}\right) = f(x_i)$. If k is the maximal number of parents of any node in the graph, we now only have to specify functions of dimension not greater than k. Hence the BBN is another concise, yet complete representation of the joint probability distribution.

The graph itself and the (conditional) independence relations that are entailed by it form the qualitative part of a BBN model. From a set of axioms described in Pearl (1988) and certain assumptions discussed later in Chapter 5, one can produce the entire set of independence relations that are implied by the BBN. An equivalent approach to determine the independence relations from the structure of a BBN is using the rules of d-separation. The concept of d-separation is detailed in Section 5.4 of Chapter 5.

The quantitative part of the model consists of the conditional probability functions associated with the variables. After these functions are quantified, the BBN can be used for probabilistic inference. Inference algorithms are available for BBNs with discrete and/or Gaussian nodes and they will be discussed in the following sections. Even though, most of these algorithms are efficient for reasonably large structures, their effectiveness is sometimes overestimated. Statements like (Langseth 2007):

> Efficient algorithms for calculating arbitrary marginal distributions [...], as well as conditional distributions [...], make BNs well suited for modeling complex systems. Models containing thousands of variables are not uncommon.

without any references to support them, can create a false image about the inference algorithms in question.

We shall further discuss the details of the different types of BBNs currently in use, taking a close look at their properties and their, often overlooked and underestimated, disadvantages: at the *facts* and at the *fiction*.

1.1.1 Discrete BBNs

In discrete BBNs nodes represent discrete random variables. These models specify marginal distributions for source nodes, and conditional probability tables (CPT) for child nodes.

Consider the BBN from Figure 1.1 with discrete nodes, each node taking k values, denoted $x_i^j, i = 1, \ldots, 4, j = 1, \ldots, k$. The marginal distributions of X_1, X_2 and X_3, and the conditional distribution of X_4 have to be specified. These distributions can be retrieved from data, when available, or elicited from experts. Table 1.1 shows the CPT for node 4.

X_1	X_2	X_3	$P(X_4 = x_4^1 \mid X_1, X_2, X_3)$	$P(X_4 = x_4^2 \mid X_1, X_2, X_3)$	...	$P(X_4 = x_4^k \mid X_1, X_2, X_3)$
x_1^1	x_2^1	x_3^1	?	?	...	?
x_1^1	x_2^1	x_3^2	?	?	...	?
...	...	...	...	...	...	...
x_1^k	x_2^k	x_3^k	?	?	...	?

Table 1.1: *Conditional probability table for X_4*

The above table contains k^4 entries. In the case of binary variables, 16 values have to be specified in a consistent manner. In absence of data, structured expert judgment should be the choice for quantifying this input. Nevertheless there are modellers who provide assessments of uncertainty themselves, and others who agree with this practice (Charniak 1991).

> [...] the skeptic might still wonder how the numbers that are still required are, in fact, obtained. In all the examples described previously, they are made up. Naturally, nobody actually makes this statement. What one really says is that they are elicited from an expert who subjectively assesses them. This statement sounds a lot better, but there is really nothing wrong with making up numbers. For one thing, experts are fairly good at it.

If the variables that form the BBN from Figure 1.1 take 10 possible values each, then the above table contains 10.000 entries, i.e. 10.000 conditional probabilities must be acquired and maintained. This would be a tremendous burden for an expert to *subjectively assess them.* A typical example of how things can go wrong in modelling complex problems with discrete BBNs is Edwards (1998).

After quantification, BBNs are used to answer probabilistic queries about the variables involved, i.e. for inference. The network can be used to update the knowledge of the state of a subset of variables when other variables (the evidence variables) are observed. There are two types of algorithms for inference: exact algorithms and approximation algorithms. In surveys of these algorithms, referring to the nature of variables from a BBN, one can find statements of the following type (Guo and Hsu 2002):

> These random variables can be either continuous or discrete. For simplicity, in this paper we shall only consider discrete ones.

This can be misleading for more than one reason. First of all the continuous variables are restricted to the normal distribution. Moreover, most of the exact algorithms were designed for discrete BBNs, and only some of them were extended to BBNs with discrete and Gaussian nodes. The latter will be discussed in the next section. The approximation algorithms are more useful for large, complex discrete structures (when exact inference algorithms are very slow), and for Gaussian structures.

Among the exact inference methods we mention *variable elimination* (Zhang and Poole 1994). The idea of this method is to use the factored representation of the joint probability distribution to do marginalisation efficiently. Irrelevant terms will be summed out (marginalised). The elimination order of the variables is not unique. The complexity of this algorithm can be measured by the number of multiplications and summations it performs. Choosing an elimination order to minimize this is NP-hard (Murphy 2002).

An alternative to variable elimination is dynamic programming, used to compute several marginals at the same time without the redundant computations that would be performed if variable elimination would be used repeatedly.

If the BBN does not have undirected cycles, a *local message passing algorithm* can be used (Pearl 1988). If it has undirected cycles, the most common approach is to convert the BBN into a tree, by clustering sets of nodes, to form a junction tree[3]. Then a local message passing algorithm is used on this tree. A variant of this method, designed for undirected models is presented in Cowell et al. (1999). The running time of this algorithms is exponential in the size of the largest cluster of nodes (Murphy 2002).

The alternative are approximation algorithms, like *variational methods*, *Monte Carlo methods*, *bounded cutset conditioning*, or *parametric approximation methods*. For details about this methods we refer to Jordan et al. (1999), Jaakkola et al. (1999), MacKay (1999), and Murphy (2002).

Except the fact that inference for large and complex discrete models can be slow, discrete BBNs suffer other serious disadvantages[4]:

- Applications involving high complexity in data-sparse environments are severely limited by the excessive assessment burden which leads to rapid, informal and indefensible quantification. This assessment burden can only be reduced by a drastic discretization of the nodes, or simplification of the model.

- The marginal distributions can often be retrieved from data, but not the

[3]Given a graph that has no chordless cycles (i.e. a triangulated graph), a junction tree is constructed by forming a maximal spanning tree from the cliques in the graph. A clique is a subgraph in which every vertex is connected to every other vertex in the subgraph.

[4]first of which was touched upon earlier in this section.

full interactions between children and parent nodes. These marginal distributions often represent the most important information driving the model; dependence information is often less important. Thus the construction of conditional probability tables should not molest any available data input. Rough discretization of course does exactly that.

- Discrete BBNs take marginal distributions only for source nodes, marginals for other nodes are computed from the conditional probability tables. When these marginals are available from data, this imposes difficult constraints on the conditional probabilities. Thus in quantification with expert judgment, it would be impractical to configure the elicitation such that the experts would comply with the marginals.

- Whereas BBNs are very flexible with respect to recalculation and updating, they are not flexible with respect to changes in modelling: if we add one parent node, then we must re-do all previous quantification for the children of this node.

Some of the drawbacks listed above are also mentioned in Cowell et al. (1999).

1.1.2 Gaussian and Discrete-Gaussian BBNs

If the nodes of a BBN correspond to variables that follow a joint normal distribution, we talk of Gaussian BBNs (or normal BBNs) (Pearl 1988; Shachter and Kenley 1989).

Continuous BBNs developed for joint normal variables interpret *influence* of the parents on a child as partial regression coefficients when the child is regressed on the parents. They require means, conditional variances and partial regression coefficients which can be specified in an algebraically independent manner (Shachter and Kenley 1989).

Let let $X = (X_1, ..., X_n)$ have a multivariate normal distribution. For Gaussian BBNs the conditional probability functions associated with the variables are of the form:

$$f\left(X_i|X_{Pa(i)}\right) \sim \mathcal{N}\left(\mu_i + \sum_{j\in Pa(i)} b_{ij}(X_j - \mu_j);\ \nu_i\right),$$

where $\mu = (\mu_1, ..., \mu_n)$ is the mean vector, $\nu = (\nu_1, ..., \nu_n)$ is a vector of conditional variances and b_{ij} are linear coefficients that can be thought of as partial regression coefficients $b_{ij} = b_{ij;Pa(i)\setminus j}$.
Continuous BBNs as above are much easier to construct than their discrete counterparts if the joint distribution is indeed normal. In absence of data, for each arc a conditional regression coefficient must be assessed. This is the answer to a question of the following type: *"Suppose that one parent variable were moved up by One Normal Unit, by how many Normal Units would you expect the child to*

move?"

One can also construct a discrete-continuous model (Cowell et al. 1999) in which continuous nodes can have discrete parents but not discrete children[5] and the conditional distribution of the continuous variables given the discrete variables is multivariate normal.

As mentioned in the previous section, some exact inference algorithm for discrete BBNs, were extended for BBNs with conditional normal distributions (Pearl 1988 and Cowell et al. 1999). Other algorithms were introduced in Lauritzen (1992) and Lauritzen and Jensen (2001). The former proved numerically unstable, and the latter requires evaluations of matrix generalized inverses and recursive combinations of potentials[6], which makes it complicated (Cowell 2005). Another algorithm is presented in Cowell (2005). The computations are performed on an elimination tree[7], rather than on a junction tree.

The price of the Gaussian and discrete-Gaussian BBNs is the restriction to the joint normal distribution, and, in the absence of data, to experts who can asses partial regression coefficients and (by assumption) constant conditional variances. If the normality assumption does not hold, then:

- The individual variables must be transformed to normals (requiring of course the marginal distributions);
- The conditional variance in *Normal Units* must be constant;
- The partial regression coefficients apply to the normal units of the transformed variables, not to the original units. This places a heavy burden on any expert elicitation;
- If a parent node is added or removed, after quantification, then the previously assessed partial regression coefficients must be re-assessed. This reflects the fact that partial regression coefficients depend on the set of regressors.

Hence, circumventing the restriction to joint normality is primarily of theoretical interest.

1.1.3 Non-parametric BBNs

Until recently, there where two ways of dealing with continuous BBNs. One was to discretize the continuous variables and work with the coresponding discrete model,

[5] Theoretically there is no need for such a restriction. However in applications, if this restriction is violated, some conditional marginals become mixtures of normals and this extension is technically demanding (Cowell et al. 1999).

[6] A potential is associated with each clique; it is a non-negative function on the realizations of that clique.

[7] An elimination tree is similar to a junction tree, in that it is a tree structure, but with the node set being a subset of the complete subgraphs of a chordal graph (rather than the set of cliques).

and the other was to assume joint normality. Both these methods have serious drawbacks, as discussed in the previous sections. In Kurowicka and Cooke (2004) the authors introduced an approach to continuous BBNs using vines (Cooke 1997; Bedford and Cooke 2002) together with copulae that represent (conditional) independence as zero (conditional) rank correlation. Copulae and vines are discussed in the next section. Suffice to say here that a copula is a distribution on the unit square, with uniform marginal distributions; and vines are graphical models that represent multivariate distributions using bivariate and conditional bivariate pieces. Moreover there is a close relationship between vines and BBNs.

In the procedure proposed in Kurowicka and Cooke (2004), nodes are associated with arbitrary continuous invertible distributions and arcs with (conditional) rank correlations, which are realized by the chosen copula. No joint distribution is assumed, which makes the BBN non-parametric. In order to quantify BBNs using this approach, one needs to specify all one dimensional marginal distributions and a number of (conditional) rank correlations equal to the number of arcs in the BBN. These assignments together with the BBN structure, the choice of the copula, and the marginals uniquely determine the joint distribution. The (conditional) rank correlations assigned to the edges of a BBN are algebraically independent. The dependence structure is meaningful for any such quantification, and need not be revised if the univariate distributions are changed. Moreover if a parent node is added or removed, after quantification, then the previously assessed (conditional) rank correlations need not be re-assessed.

One way of stipulating a joint distribution is by sampling it. The sampling algorithm for BBNs, using vines, is fully described in Chapter 2. The sampling procedure works with arbitrary conditional copulae. Thus it can happen that variables X, and Y are positively correlated when variable Z takes low values, but are negatively correlated when Z is high. This behaviour indicates that it would be appropriate to use non-constant conditional copulae (hence non-constant conditional correlations), but the use of such copulae would significantly complicate the Monte Carlo sampling and the assessment. We will therefore restrict our study to *constant* conditional rank correlations.

Conditional rank correlations are not elicited directly or estimated from data directly. Rather, given a copula, these can be obtained from conditional exceedance probabilities. Thus suppose node A has parents B and C. According to the protocol described in Section 2.1, we need the rank correlation r_{AB} and the conditional rank correlation $r_{AC|B}$. We extract these from answers to the following two questions (Morales et al. 2007):

- *"Suppose that B was observed to be above its median, what is the probability that A is also above its median?"*

- *"Suppose that B and C were both observed to be above their medians, what is the probability that A is also above its median?"*

The relationship between the conditional exceedence probabilities and the (conditional) rank correlations depends on the choice of copula. Moreover, the answer to the second question is constrained by the expert's answers to previous question. Hence bounds for the conditional probability of exceedance (at each step of the elicitation) have to be computed. Other elicitation procedures are also developed. For details we refer to Morales et al. (2007).

The conditional rank correlations, obtained in the way described above, can be realized using any copula that represents (conditional) independence as zero (conditional) rank correlation.

The copula-vine modelling approach is general and allows defensible quantification methods, but it comes at the price that these BBNs must be evaluated by Monte Carlo simulation. Updating such a BBN requires re-sampling the whole structure every time evidence becomes available. Moreover, there are situations in which sampling large complex structures only once can still involve very time consuming numerical calculations.

1.2 Copulae & Vines

We introduce notations and terminology needed throughout the subsequent chapters. The emphasis is on *copulae* and *vines*. Most of the concepts presented here can be found in Kurowicka and Cooke (2006b). If not, alternative references are given.

Definition 1.2.1. *The* copula *of two continuous random variables X and Y is the joint distribution of $F_X(X)$ and $F_Y(Y)$, where F_X, F_Y are the cumulative distribution functions of X, Y respectively. The copula of (X, Y) is a distribution on $[0,1]^2 = \boldsymbol{I}^2$ with uniform marginal distributions.*

An overview of copulae can be found in Nelsen (1999),or Joe (1997). Here, we only list a small number of families of copulae that will be used in this thesis.

1. *Independence copula*

$$\Pi(u, v) = uv, \quad (u, v) \in \mathbf{I}^2.$$

2. *Fréchet upper bound copula*

$$M(u, v) = \min(u, v), \quad (u, v) \in \mathbf{I}^2.$$

3. *Fréchet lower bound copula*

$$W(u, v) = \max(0, u + v - 1), \quad (u, v) \in \mathbf{I}^2.$$

4. *Normal copula*

   If Φ_ρ is the bivariate normal cumulative distribution function with product moment correlation ρ and Φ^{-1} the inverse of the standard univariate normal distribution function then:

$$C_\rho(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right), \quad (u,v) \in \mathbf{I}^2.$$

5. *Frank's copula*(Frank 1979)

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right), \quad (u,v) \in \mathbf{I}^2,\ \theta \in (-\infty,\infty).$$

   When $\theta \to \infty$ ($\theta \to -\infty$) then Frank's copula corresponds to M (W). The limit $\theta \to 0$ yields the independence copula Π.

6. *Mardia copula*

$$C_\theta(u,v) = \frac{\theta^2(1+\theta)}{2}M(u,v) + (1-\theta^2)\Pi(u,v) + \frac{\theta^2(1-\theta)}{2}W(u,v),$$

   where $(u,v) \in \mathbf{I}^2,\ \theta \in [-1,1]$.

For every copula C and every $(u,v) \in \mathbf{I}^2$,

$$W(u,v) \le C(u,v) \le M(u,v).$$

The above inequalities suggest a partial order on the set of copulae.

Definition 1.2.2. *If C_1 and C_2 are copulae, we say that C_1 **is smaller than** C_2 and write $C_1 \prec C_2$ if $C_1(u,v) \le C_2(u,v)$ for all $(u,v) \in \mathbf{I}^2$.*

However, there are families of copulae which are totally ordered.

Definition 1.2.3. *We call a totally ordered parametric family $\{C_\theta\}$ of copulae **positively ordered** if $C_\alpha \prec C_\beta$ whenever $\alpha \le \beta$.*

As examples of positively ordered copulae we mention Frank's copula, and the normal copula. The Mardia copula on the other hand is an unordered copula (Nelsen 1999).

A useful property of a copula is that of representing independence as zero correlation. Such copula is said to have the *zero independence property.*

We shall now move on to define the graphical models called *vines.*

Vines were introduced in Cooke (1997) and Bedford and Cooke (2002). A vine on n variables is a nested set of trees. The edges of the $j^{\underline{th}}$ tree are the nodes of the $(j+1)^{\underline{th}}$ tree. A *regular* vine on n variables is a vine in which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node. More formally:

Definition 1.2.4. *$\mathcal{V}$ is called a* regular vine *on n elements if:*

1. *$\mathcal{V} = (T_1, \ldots, T_{n-1})$;*
2. *T_1 is a tree with nodes $N_1 = \{1, \ldots, n\}$, and edges E_1 and for $i = 2, \ldots, n-1$ T_i is a tree with nodes $N_i = E_{i-1}$;*
3. *For $i = 2, \ldots, n-1$, $a, b \in E_i$, $\#a \triangle b = 2$, where $\triangle$ denotes the symmetric difference. In other words if a and b are nodes of T_i connected by an edge in T_i, where $a = \{a_1, a_2\}$, $b = \{b_1, b_2\}$, then exactly one of the a_i equals one of the b_i*

We will distinguish two particular regular vines. A regular vine is called a:

- ***D-vine*** if each node in T_1 has the degree at most 2 (see Figure 1.2 (left));
- ***C-vine*** if each tree T_i has a unique node of degree $n - i$. The node with maximal degree in T_1 is called the root (see Figure 1.2 (right)).

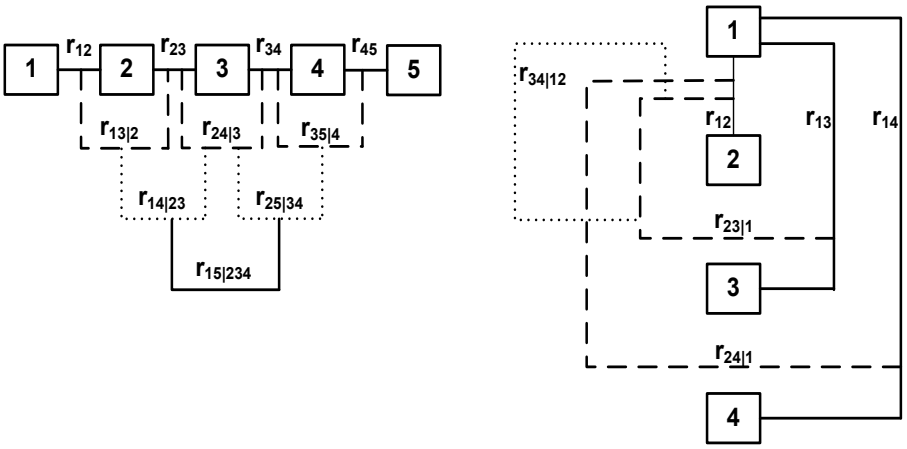


Figure 1.2: *A D-vine on 5 variables (left) and a C-vine (right) on 4 variables showing the (conditional) rank correlations associated with the edges.*

For each edge of the vine we distinguish a *constraint*, a *conditioning*, and a *conditioned* set. Variables reachable from an edge via the membership relation, form its constraint set. If two edges are joined by an edge in the next tree the intersection and symmetric difference of their constraint sets give the conditioning and conditioned sets, respectively.

Each edge of a regular vine may be associated with a constant (conditional)

rank correlation[8] which can be arbitrarily chosen in the interval $[-1, 1]$ (see Figure 1.2). Using a copula to realize these (conditional) rank correlations, a joint distribution satisfying the copula-vine specification can be constructed and it will always be consistent. For rigorous definitions and proofs we refer to Kurowicka and Cooke (2006b).

Each vine[9] edge may also be associated with a partial correlation. Partial correlations can be defined in terms of partial regression coefficients. Let us consider variables X_i with zero mean and standard deviations σ_i, $i = 1, ..., n$. Let the numbers $b_{12;3,...,n}$,...,$b_{1n;2,...,n-1}$ minimise:

$$E\left((X_1 - b_{12;3,...,n}X_2 - ... - b_{1n;2,...,n-1}Xn)^2\right).$$

Definition 1.2.5. *The partial correlation of X_1 and X_2 based on X_3,..., X_n is:*

$$\rho_{12;3,...,n} = sgn(b_{12;3,...,n})(b_{12;3,...,n}b_{21;3,...,n})^{\frac{1}{2}}.$$

Equivalently we could define the partial correlation as:

$$\rho_{12;3,...,n} = -\frac{C_{12}}{\sqrt{C_{11}C_{22}}},$$

where C_{ij} denotes the $(i, j)^{th}$ cofactor of the correlation matrix.
The partial correlation $\rho_{12;3,...,n}$ can be interpreted as the correlation between the orthogonal projections of X_1 and X_2 on the plane orthogonal to the space spanned by X_3,...,X_n.

Partial correlations can be computed from correlations with the following recursive formula (Yule and Kendall 1965):

$$\rho_{12;3,...,n} = \frac{\rho_{12;4,...,n} - \rho_{13;4,...,n} \cdot \rho_{23;4,...,n}}{((1 - \rho^2_{13;4,...,n}) \cdot (1 - \rho^2_{23;4,...,n}))^{\frac{1}{2}}}. \tag{1.2.1}$$

A *complete partial correlation vine specification* is a regular vine with a partial correlation specified for each edge. A partial correlation vine specification does not uniquely specify a joint distribution[10], but there is a joint distribution satisfying the specified information (Bedford and Cooke 2002). For example a joint normal distribution.

A *complete normal partial correlation specification* is a special case of a regular vine specification. The following theorem shows how the notion of a regular vine can be used to construct a joint normal distribution (Bedford and Cooke 2002).

Theorem 1.2.1. *Given any complete partial correlation vine specification there is a unique joint normally distributed random vector $(X_1, \ldots, X_n)$ satisfying all partial correlation specifications.*

[8] When we speak of rank correlation we refer to the Spearman's rank correlation. We use r to denote it. The letter ρ is used to represent the product moment correlation.

[9] Further in this thesis, whenever we speak of vines we mean regular vines.

[10] Moreover a given set of marginal distributions may not be consistent with a given set of partial correlations.

The notion of *normal vines* arises when $X_1, \ldots, X_n$ have a joint normal distribution, and the edges of a regular vine on n nodes are assigned the partial correlations of this distribution. Another important result from Bedford and Cooke (2002) is that each partial correlation vine specification uniquely determines the correlation matrix, even without the assumption of joint normality.

Theorem 1.2.2. *For any regular vine on n elements there is a one to one correspondence between the set of $n \times n$ positive definite correlation matrices and the set of partial correlation specifications for the vine.*

The joint normal copula has a well known property inherited from the joint normal distribution namely: the zero partial correlation is sufficient for conditional independence[11]. This follows from two facts: for the joint normal variables the partial correlation is equal to the conditional correlation and zero conditional correlation means conditional independence. Moreover, the relationship between the product moment correlation (ρ) and the rank correlation (r) for joint normal, is given by the Pearson's transformation, and it translates these properties to normal copula.

Proposition 1.2.1. *(Pearson 1907) Let (X, Y) be a random vector with the joint normal distribution, then:*

$$\rho(X, Y) = 2\sin(\tfrac{\pi}{6} \cdot r(X, Y)).$$

The property of vines that plays a crucial role in model inference is given in the next theorem(Kurowicka and Cooke 2006a).

Theorem 1.2.3. *Let D be the determinant of the correlation matrix of variables $X_1, \cdots, X_n$, with $D > 0$. For any partial correlation vine*

$$D = \prod_{e \in E(\mathcal{V})} \big(1 - \rho^2_{e_1, e_2; D_e}\big),$$

where $E(\mathcal{V})$ is the set of edges of the vine $\mathcal{V}$, D_e denotes the conditioning set associated with edge e, and $\{e_1, e_2\}$ is the conditioned set of e.

Vines are actually a way of factorising the determinant of the correlation matrix. The key notion in deriving the equation from Theorem 1.2.3 is *multiple correlation.*

Definition 1.2.6. *The multiple correlation $R_{1:2,\ldots,n}$ of variables 1 with respect to $2, ..., n$ is:*

$$1 - R^2_{1:2,\ldots,n} = \frac{D}{C_{11}},$$

where D is the determinant, and C_{11} is the (1,1) cofactor of the correlation matrix C.

[11] In general, conditional independence is neither necessary, nor sufficient for zero partial correlation (Kurowicka 2001).

The multiple correlation $R_{1:2,...,n}$ of variables 1 with respect to $2, ..., n$ is the correlation between 1 and the best linear predictor of 1 based on $2, ..., n$. It is easy to show that (Kurowicka and Cooke 2006b):

$$D = \left(1 - R^2_{1:2,...,n}\right)\left(1 - R^2_{2:3,...,n}\right)...\left(1 - R^2_{n-1:n}\right). \tag{1.2.2}$$

In (Kendall and Stuart 1961) it is shown that $R_{1:2,...,n}$ is non negative and satisfies:

$$1 - R^2_{1:2,...,n} = (1 - \rho^2_{1n})(1 - \rho^2_{1n-1;n})(1 - \rho^2_{1n-2;n-1,n})...(1 - \rho^2_{12;3,...,n}).$$

The concept of multiple correlation and its relationship with partial correlations will be required later, in Chapter 5, when proving a similar property for the partial correlation specification for BBNs.

1.3 Aim of Research & Reading Guide

The starting point of this research is the approach from Kurowicka and Cooke (2004). This method applies to non-parametric continuous BBNs. It is a general and flexible approach. Nevertheless there are BBN structures for which sampling even once might be very complicated and time consuming under certain conditions. The first objective of our research is to overcome this problem and develop further an algorithm such that it is fast in any circumstances. Often, real life problems involve a large number of variables, connected in complex ways, hence the algorithm should cope with these situations. Another objective is to extend the theory for non-parametric continuous BBNs to include ordinal discrete random variables. In the last part of our research, we use BBNs as tools for mining ordinal multivariate data. We aim to develop an algorithm for learning the structure of a BBN from an ordinal data set.

The objectives formulated above are dealt with in 5 chapters of the thesis. Chapter 2 reviews the details of non-parametric BBNs using the copula-vine modelling approach and introduces two new methods. The first one is a hybrid approach, which consists of combining the reduced assessment burden and modelling flexibility of the continuous BBNs with the fast updating algorithms of discrete BBNs. This is done, using vine sampling together with existing discrete BBNs software. The drawbacks of this method are discussed, and a second method is introduced. A new sampling protocol based on the normal copula is proposed. Normal vines are used to realize the dependence structure specified via (conditional) rank correlations on the continuous BBN.

In order to extend this approach to include ordinal discrete random variables we need to study the concept of rank correlation between two such variables. In contrast with the continuous case, 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. Chapter 3 presents a generalisation of the population version of Spearman's rank correlation for the case of ordinal discrete random variables.

Discrete univariate distributions can be obtained as monotone transforms of uniform variables. A class of discrete bivariate distributions can be constructed by specifying the marginal distributions and a copula. The rank correlation coefficient of the discrete variables depends on not only the copula, but also the marginal distributions. An analytical description of this dependence is derived and discussed in case of different copulae and different marginal distributions.

In Chapter 4 we present two large ongoing projects in which mixed non-parametric continuous & discrete BBNs are the tool used in the analysis.

Chapter 5 is concerned with non-parametric BBNs from a completely different point of view, namely as a tool for mining ordinal multivariate data. We propose a method for learning a BBN from data. The main advantage of this method is that it can handle a large number of continuous variables, without making any assumption about their marginal distributions, in a very fast manner. Once we have learned the BBN from data, we can further use it for prediction or diagnosis by employing the methods described in the previous chapters. We illustrate the method proposed using a database of pollutants emissions and fine particulate concentrations.

In Chapter 6 the most important results of this work are summarised and conclusions are formulated. Finally, a short software description, and some technical details are given in Chapter 7.

Chapter 2

Methods for Quantifying and Analyzing BBNs[1]

Since BBNs have become a popular tool for specifying high dimensional probabilistic models, commercial tools with an advanced graphical user interface that support their construction and inference are available. Thus, building and working with BBNs is very efficient as long as one is not forced to quantify complex BBNs. A high assessment burden of discrete BBNs is often caused by the discretization of continuous variables. An alternative to the discretization of continuous variables or the assumption of normality is the *copula-vine approach* to continuous BBNs. The details of this approach are discussed in the beginning of this chapter. The approach is quite general and allows traceable and defensible quantification methods, but it comes at a price: the BBNs must be evaluated by Monte Carlo simulation. Updating such a BBN requires re-sampling the whole structure. The advantages of fast updating algorithms for discrete BBNs are decisive. A hybrid method advanced in Section 2.2 samples the continuous BBN once, and then discretizes this so as to enable fast updating. This combines the reduced assessment burden and modelling flexibility of the continuous BBNs with the fast updating algorithms of discrete BBNs.

Sampling large complex structures only once can still involve time consuming numerical calculations. Therefore a new sampling protocol is developed (Section 2.3). Given that the conditional copulae do not depend on conditioning variables, there are great advantages to using the joint normal copulae, hence this new protocol is based on normal vines.

The last section of this chapter describes a very important feature of the normal copula vine method, namely that conditioning can be done analytically.

[1]This chapter is based on the paper Hanea et al. (2006), "Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets", published in Quality and Reliability Engineering International, 22(6).

2.1 Continuous BBNs & Vines

The nodes of a non-parametric continuous BBN represent continuous univariate random variables. The arcs are associated with (conditional) parent-child rank correlations. We assume throughout this chapter that all univariate distributions have been transformed to uniform distributions on $(0, 1)$. Any copula with invertible conditional cumulative distribution function may be used as long as it represents (conditional) independence as zero (conditional) correlation. We note that quantifying BBNs in this way requires assessing all (continuous, invertible) one dimensional marginal distributions. One can assign (conditional) rank correlations to the arcs of a BBN according to the protocol presented in Kurowicka and Cooke (2004). The conditional rank correlations need not be constant, although they are taken to be constant in the following examples. In contrast, in Section 2.3, where we introduce normal vines, the conditional rank correlations must be constant. We will illustrate the protocol for assigning (conditional) rank correlations to the arcs of a BBN with an example.

Example 2.1.1. *Let us consider the undirected cycle on* 4 *variables from Figure 2.1.*

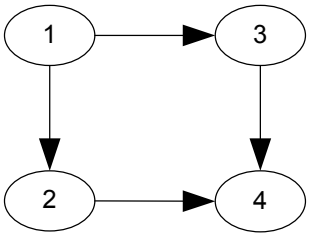


Figure 2.1: *BBN with 4 nodes and 4 arcs.*

There are two sampling orders for this structure: 1, 2, 3, 4, or 1, 3, 2, 4. Let us choose 1, 2, 3, 4. The factorization of the joint distribution is:

$$P(1)P(2|1)P(3|1\underline{2})P(4|23\underline{1}). \tag{2.1.1}$$

The underscored nodes in each conditioning set are the non-parents of the conditioned variable. Thus they are not necessary in sampling the conditioned variable. This uses some of the conditional independence relations in the belief net. If they would be omitted from the conditioning set, the factorisation (2.1.1) would coincide with the factorisation (1.1.2). To each arc of the BBN we will assign a parent-child rank correlation. The correlation between the child and its first parent[2] *will be an unconditional rank correlation, and the correlations between the*

[2]The parents of each variable can be ordered in a non-unique way.

child and its next parents (in the ordering) will be conditioned on the values of the previous parents. Hence, one set of (conditional) rank correlations that can be assigned to the edges of the BBN from Figure 2.1 are[3]*:* $\{r_{21}, r_{31}, r_{42}, r_{43|2}\}$. *For each term* i *(*$i = 1, \ldots, 4$*) of the factorization 2.1.1, a D-vine on* i *variables is built. This D-vine is denoted by* $\mathcal{D}^i$ *and it contains: the variable* i*, the non-underscored variables, and the underscored ones, in this order. Figure 2.2 shows the D-vines built for variables 2, 3, 4.*

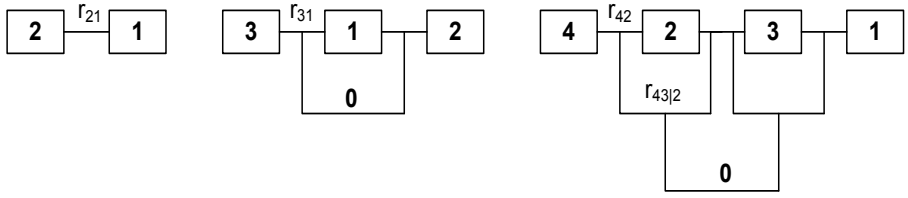


Figure 2.2: $\mathcal{D}^2, \mathcal{D}^3, \mathcal{D}^4$ *for Example 2.1.1.*

Building the D-vines is not a necessary step in specifying the rank correlations[4], but it is essential in proving the main result for continuous BBNs. In order to formulate this result, we need a more general setting. For a BBN on n variables the factorization of the joint distribution in the standard way (following the sampling order $1, \ldots, n$) is:

$$P(1, \ldots, n) = P(1)P(2|1)P(3|2, 1) \ldots P(n|n-1, \ldots, 1). \tag{2.1.2}$$

In this factorization, we will underscore the nodes from each conditioning set, which are not parents of the conditioned variable. For each term i with parents (non-underscored variables) $i_1 ... i_{p(i)}$ in equation (2.1.2), we associate the arc $i_{p(i)-k} \longrightarrow i$ with the conditional rank correlation:

$$\begin{cases} r(i, i_{p(i)}), & k = 0 \\ r(i, i_{p(i)-k} | i_{p(i)}, ..., i_{p(i)-k+1}), & 1 \leq k \leq p(i) - 1. \end{cases} \tag{2.1.3}$$

The assignment is vacuous if $\{i_1 ... i_{p(i)}\} = \emptyset$. Assigning (conditional) rank correlations for $i = 1, ..., n$, every arc in the BBN is assigned a (conditional) rank correlation between parent and child.

The following theorem is crucial for the copula vine approach to non-parametric continuous BBNs. It shows that these assignments uniquely determine the joint distribution and are algebraically independent.

[3]One could as well specify $\{r_{21}, r_{31}, r_{43}, r_{42|3}\}$ instead.

[4]These are assigned directly to the arcs of the BBN. Each arc is associated with a (conditional) parent-child rank correlation.

Theorem 2.1.1. *Given:*

1. *a directed acyclic graph with n nodes specifying conditional independence relationships in a BBN;*
2. *n variables, assigned to the nodes, with continuous invertible distribution functions;*
3. *the specification (2.1.3), $i = 1, ..., n$ of conditional rank correlations on the arcs of the BBN;*
4. *a copula realizing all correlations $[-1, 1]$ for which correlation 0 entails independence;*

the joint distribution of the n variables is uniquely determined. This joint distribution satisfies the characteristic factorization (2.1.2) and the conditional rank correlations in (2.1.3) are algebraically independent.

Proof. Given that all univariate distributions are known, continuous, invertible functions, one can use them to transform each variable to a uniform on $(0, 1)$. Hence, we can assume, without any loss of generality, that all univariate distributions are uniform distributions on $(0, 1)$.

The first term in (2.1.3) is determined vacuously. We assume the joint distribution for $\{1, ..., i-1\}$ has been determined. The $i^{\underline{th}}$ term of the factorization (2.1.2) involves $i-1$ conditional variables, of which $\{i_{p(i)+1}, ..., i_{i-1}\}$ are conditionally independent of i given $\{i_1, ..., i_{p(i)}\}$. We assign:

$$r(i, i_j | i_1, ...i_{p(i)}) = 0;\ i_{p(i)} < i_j \leq i - 1. \tag{2.1.4}$$

Then the conditional rank correlations (2.1.3) and (2.1.4) are exactly those on $\mathcal{D}^i$ involving variable i. The other conditional bivariate distributions on $\mathcal{D}^i$ are already determined. It follows that the distribution on $\{1, ..i\}$ is uniquely determined. Since zero conditional rank correlation implies conditional independence,

$$P(1, ...i) = P(i|1...i-1)P(1, ..., i-1) = P(i|i_1...i_{p(i)})P(1, ..., i-1).$$

from which it follows that the factorization (2.1.2) holds. The fact that the (conditional) rank correlations are algebraically independent follows immediately from the same property of the rank correlation specification on a regular vine (Kurowicka and Cooke 2006b). □

The (conditional) rank correlations and the marginal distributions needed in order to specify the joint distributions represented by the BBN, can be retrieved from data, if available or elicited from experts (Morales et al. 2007).

After specifying the joint distribution, we will now show how to sample it. In order to sample a BBN structure we will use the procedures for vines. We can sample X_i using the sampling procedure for the vine $\mathcal{D}^i$. When using vines to

sample a continuous BBN, it is not in general possible to keep the same order of variables in successive D-vines. In other words, we will have to re-order the variables before constructing $\mathcal{D}^{i+1}$ and sampling X_{i+1}, and this will involve calculating some conditional distributions. We will present the sampling procedure for BBNs using the structure from Example 2.1.1. In Figure 2.2, one can notice that the D-vine for the $3^{\underline{rd}}$ variable is $\mathcal{D}^3 = D(3,1,2)$, and the order of the variables from $\mathcal{D}^4$ must be $D(4,3,2,1)$. Hence, this BBN cannot be represented as just one D-vine. An example of a BBN structure that can be represented as one single D-vine is given in Figure 1.1 from Chapter 1. Its equivalent D-vine is showed in Figure 2.3.

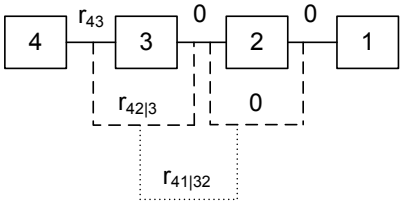


Figure 2.3: *The D-vine corresponding to the BBN from Figure 1.1.*

Let us return to the sampling procedure for the BBN structure from Example 2.1.1. We start with sampling four independent, uniform (0,1) variables, say $U_1, \ldots, U_4$.

$x_1 = u_1;$
$x_2 = F^{-1}_{r_{21};x_1}(u_2);$
$x_3 = F^{-1}_{r_{31};x_1}(F^{-1}_{r_{32|1};F_{r_{21};x_1}(x_2)}(u_3));$
$x_4 = F^{-1}_{r_{42};x_2}(F^{-1}_{r_{43|2};F_{r_{32};x_2}(x_3)}(F^{-1}_{r_{41|32};F_{r_{21|3};F_{r_{32};x_3}(x_2)}(F_{r_{31};x_3}(x_1))}(u_4)))$,

where $F_{r_{ij|k};X_i}(X_j)$ denotes the cumulative distribution function of X_j, given X_i under the conditional copula with correlation $r_{ij|k}$.

The BBN structure reads the conditional independence of X_3 and X_2 given X_1 ($r_{32|1} = 0$), and of X_4 and X_1 given X_2, X_3 ($r_{41|32} = 0$), hence:

$$F^{-1}_{r_{32|1};F_{r_{21};x_1}(x_2)}(u_3) = u_3 \text{ and } F^{-1}_{r_{41|32};F_{r_{21|3};F_{r_{32};x_3}(x_2)}(F_{r_{31};x_3}(x_1))}(u_4) = u_4.$$

Consequently, using these conditional independence properties, the sampling procedure can be simplified as:

$x_1 = u_1;$
$x_2 = F^{-1}_{r_{21};x_1}(u_2);$
$x_3 = F^{-1}_{r_{31};x_1}(u_3);$

$x_4 = F^{-1}_{r_{42};x_2}(F^{-1}_{r_{43|2};F_{r_{32};x_2}(x_3)}(u_4))$.

We shorten the notation by dropping the *"r"*'s and write $F_{j|i}(x_j)$ instead of $F_{r_{ij};x_i}(x_j)$. The conditional distribution $F_{3|2}(x_3)$ is not given explicitly, but it can be calculated as follows:

$$F_{3|2}(x_3) = \int_0^{x_3}\int_0^1 c_{21}(x_2,x_1)c_{31}(v,x_1)dx_1dv,$$

where c_{i1} is the density of the chosen copula with correlation r_{i1}, $i \in \{2,3\}$. We use Frank's copula to realise the (conditional) rank correlations. The reasons for this choice are: it has the zero independence property; it realizes a specified rank correlation without adding too much information to the product of the margins; its density covers the entire unit square; it has tractable functional forms for the density; conditional distribution and inverse of the conditional distribution.

For each sample, one needs to calculate the numerical value of the double integral[5]. In this case, when only one double integral needs to be evaluated, it can be easily done without excessive computational burden.

If we observe the values of some variables, the results of sampling this model - conditional on their values - are obtained either by sampling again the structure (*cumulative approach*), or by using the *density approach*. We will present both methods in short, and for details we refer to (Kurowicka and Cooke 2006b).
Let us assume we learn $X_2 = 0.85$. In the cumulative approach the sampling procedure becomes[6]:

$x_1 = F^{-1}_{1|2:x_2}(u_1)$;
$x_2 = 0.85$;
$x_3 = F^{-1}_{3|1:x_1}(u_3)$;
$x_4 = F^{-1}_{4|2:x_2}(F^{-1}_{4|32:F_{3|2}(x_3)}(u_4))$.

In the density approach, the joint density can be evaluated as follows (Bedford and Cooke 2002):

$$g(x_1,\ldots,x_4) = c_{21}(x_2,x_1)c_{31}(x_3,x_1)c_{42}(x_4,x_2)c_{43|2}(F_{4|2}(x_4),F_{3|2}(x_3)).$$

The conditionalisation is made using $x_2 = 0.85$ in the above formula and re-sampling with weights proportional to $g(x_1, 0.85, x_3, x_4)$. Whichever of the two methods is preferred, the double integral still needs to be evaluated for each sample, and for any new conditionalisation.

If the BBN is an undirected cycle of five variables, and the same sampling procedure is applied, a triple integral will have to be calculated. The bigger the cycle is, the larger the number of multiple integrals that have to be numerically evaluated. And yet, this is not the worst that can happen[7]; an example of such a

[5] All numerical results in this chapter are obtained using Matlab.

[6] Sometimes, the sampling order has to be changed in order to perform conditioning using the cumulative approach.

[7] More examples of BBN structures in which additional numerical calculations are needed are presented in Chapter 6 of Kurowicka and Cooke (2006b).

situation will be presented in Section 2.3 of this chapter.

The BBNs that resemble real life problems will often be quite large, and may well contain undirected cycles of five or more variables. Updating such a structure is done by re-sampling the network each time new evidence is obtained. In case of a large number of variables, one would have to be prepared to run the model for a few days. To overcome this limitation we would like to combine the vine approach to the continuous BBNs, with the benefits of the discrete BBNs software. This is done in the next section.

2.2 Hybrid Method

Sampling a large BBN structure every time new evidence becomes available does not seem a very good idea in terms of computational time. On the other hand, sampling it just once, and employing the easiness of use, flexibility, good visualisation, and fast updating of a commercial BBN tool, provides an elegant solution to this problem. The hybrid method proposed here can be summarised as follows:

1. Quantify nodes of a BBN as continuous univariate random variables and arcs as parent-child (conditional) rank correlations;

2. Sample this structure creating a large sample file;

3. Use this sample file (in a commercial BBN tool) to build conditional probability tables for a discretized version of the continuous BBN;

4. Use the commercial tool to visualise the network and perform fast updating for the discretized BBN.

Most often, when continuous non-parametric BBNs have to be quantified, their discretized version is used instead. A large number of states should be used for each node, in order for the quantification to be useful. This leads to huge conditional probability tables that must be filled in, in a consistent manner. In contrast, the $1^{\underline{st}}$ step of the hybrid method can significantly reduce the assessment burden, while preserving the interpretation of arrows as influences. Not only is the degree of realism greater in the continuous model, but also the quantification requires only the marginal distributions and a reduced number of algebraically independent (conditional) rank correlations. After quantifying the continuous model, the discretized version of the model is used. Discretizing the nodes in fairly many states will ensure preserving the dependence structure specified via (conditional) rank correlations. The conditional probability tables for the discretized version of the model are immediately constructed, by simply importing the sample file in a commercial BBN tool ($3^{\underline{rd}}$ step of the hybrid method). The main use of the BBNs is updating on the basis of newly available information. We have shown how this can be done using the copula-vine method and what its disadvantages are. This motivates the $4^{\underline{th}}$ step of the hybrid method which offers immediate updating.

There is a large variety of BBN software tools. Some of them are free (e.g. Bayda, BNT, BUGS, GeNIe) and others are commercial, although most of the latter have free versions which are restricted in various ways (Murphy 2002) . In our experience, the commercial tools have some advantages over the free ones, either from the functionality point of view, or even because of the graphical user interface which is sometimes not included in the free software. Two of the most popular commercial tools for BBNs are Hugin[8] and Netica[9]. They both provide an elegant graphical user interface and their main features are very similar (at least the features that we use in our study). We chose Netica for our further study.

In order to perform the $3^{\underline{rd}}$ step of the hybrid method, a network has to be pre-prepared in Netica. This will contain the nodes of the BBN, each discretized in a certain - not necessarily small - number of states, together with the connections. The way in which variables are discretized is a choice of the analyst. To preserve the information about the dependence structure in the sample file, a large number of discretization intervals is preferred. On the other hand, when the number of discretization intervals for each variable increases, the size of the conditional probability tables that Netica constructs from the sample file increases as well. There is a trade off between the number of discretization intervals and the size of the conditional probability tables. After a few comparisons (for particular cases) between the choices of 5, 10 and 20 discretization intervals (for each variable), one can observe that the dependence structure assigned by the experts is maintained up to a difference of order 10^{-3} in the case where the variables are discretized in 10 intervals each, and the sample file imported in Netica does not need to be of extraordinary size. Based on this result, the variables from the following examples will be discretized each in 10 intervals. Another choice that one has to make also with respect to the discretization, is the size of the discretization intervals. The variables can be discretized in equal intervals, or according to the quantiles of their distributions, or at random. The third choice is of course not very useful. After the sample file is imported in Netica, the marginal distributions can be visualized (via the option *Style/Belief bars*). If the variables are discretized in equal intervals, the shape representing each variable corresponds to the shape its real distribution. If, on the other hand, the variables are discretized according to their quantiles, Netica will show uniform marginals. We shall illustrate the method described above by means of an extensive example.

Example 2.2.1. *Flight Crew Alertness*

In Figure 2.4, the flight crew alertness model is given. A discrete form of this model was first presented in Roelen et al. (2004) and an adapted version of it was discussed in Kurowicka and Cooke (2004). In the original model all chance nodes

[8] A light version of Hugin can be downloaded from www.hugin.com
[9] A light version of Netica can be downloaded from www.norsys.com

were discretized to take one of two values OK or NotOK. The names of nodes have been altered to indicate how, with greater realism, these can be modelled as continuous variables. Alertness is measured by performance on a simple tracking test programmed on a palmtop computer. Crew members did this test during breaks in-flight under various conditions. The results are scored on an increasing scale and can be modelled as a continuous variable. The alertness of the crew is influenced by a number of factors like: how much time the crew slept before the flight, the recent work load, the number of hours flown up until this moment in the flight (flight duty period), pre-flight fitness, etc. Figure 2.4 resembles the latest version of the model.

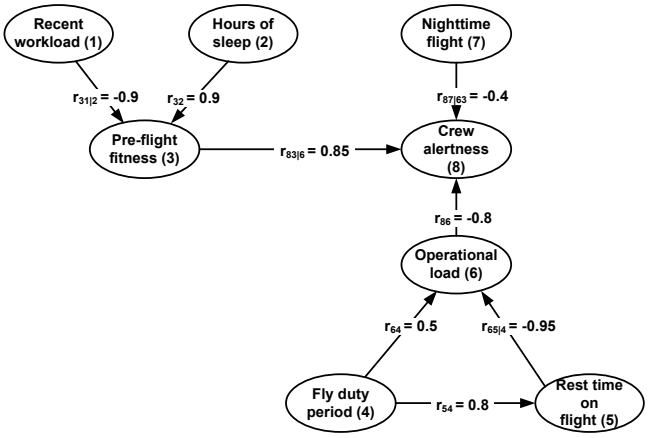


Figure 2.4: *Flight crew alertness model.*

In order to use the hybrid method described in the beginning of this section, continuous distributions for each node and (conditional) rank correlations for each arc must be gathered from existing data or expert judgement. The distribution functions are used to transform the variables to uniforms on $(0, 1)$. The (conditional) rank correlations assigned to each arc of the BBN are chosen by the authors of Kurowicka and Cooke (2004) for illustrative purposes. The marginal distributions are chosen to be uniforms on $(0, 1)$. For simplicity, we assign a number to each variable (see Figure 2.4). We choose the sampling order: 1, 2, 3, 4, 5, 6, 7, 8. The sampling procedure uses Frank's copula, and does not require any additional calculations:

$x_1 = u_1$;
$x_2 = u_2$;
$x_3 = F^{-1}_{3|2:x_2}(F^{-1}_{3|21:x_1}(u_3))$;
$x_4 = u_4$;

$x_5 = F^{-1}_{5|4:x_4}(u_5);$
$x_6 = F^{-1}_{6|4:x_4}(F^{-1}_{6|54:F_{5|4}(x_5)}(u_6));$
$x_7 = u_7;$
$x_8 = F^{-1}_{8|6:x_6}(F^{-1}_{8|63:x_3}(F^{-1}_{8|763:x_7}(u_8))).$

Figure 2.5 shows the BBN from example 2.4, modelled in Netica. The variables are uniform on the $(0, 1)$ interval, and each is discretized in 10 states. Each of these states consists in an interval, rather than a single value. A case file containing $8 \cdot 10^5$ samples, obtained using the sampling procedure described, was imported in Netica. This automatically creates the conditional probability tables.

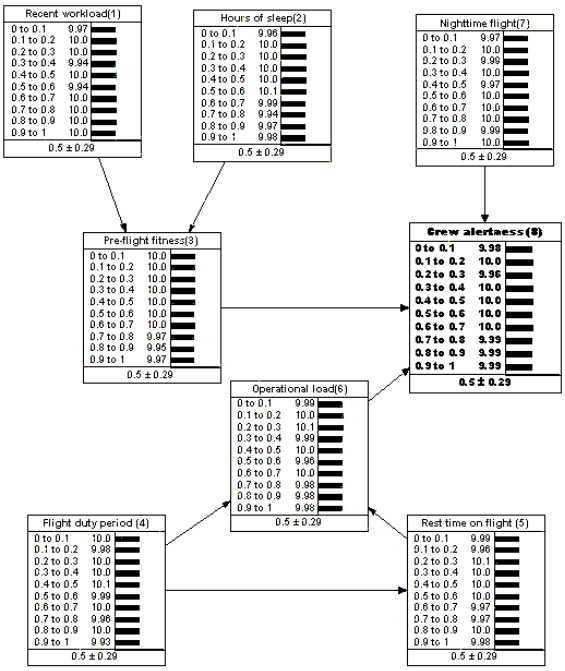


Figure 2.5: *Flight crew alertness model with histograms in Netica.*

The quantification of the discretized BBN would require 12140 probabilities, whereas the quantification with continuous nodes requires only 8 algebraically independent (conditional) rank correlations and 8 marginal distributions.

The main use of BBNs in decision support is updating on the basis of possible observations. Let us suppose that we have some information about how much the crew slept before the flight and about the flight duty period of the crew. Figures

2.6 and 2.7 present the distribution of the crew alertness in the situation when the crew's hours of sleep are between the $20^{\underline{th}}$ and the $30^{\underline{th}}$ percentiles (the crew did not have enough sleep) and the flight duty period is between the $80^{\underline{th}}$ and $90^{\underline{th}}$ percentiles (the flight duty period is long).

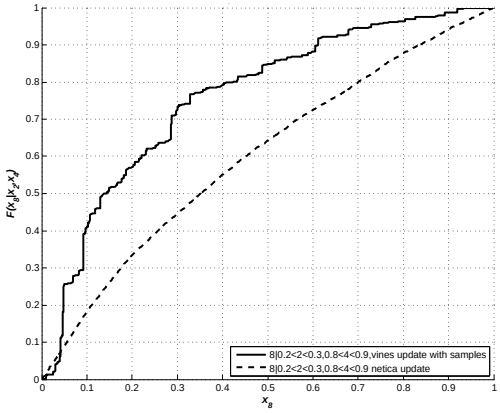


Figure 2.6: *Distribution of $X_8|X_2, X_4$. Comparison of updating results in vines and Netica using 10^4 samples.*

The conditional distribution of the Flight crew alertness(8) from Figures 2.6 and 2.7 is obtained in two ways:

- using the vines-Netica updating;
- using the vines updating with the density approach.

After the sample file is imported in Netica, we condition on Hours of sleep $\in [0.2, 0.3]$ and Fly duty period $\in [0.8, 0.9]$. We can use Netica to generate samples from the conditional distribution of Crew alertness. Even though Crew alertness appears as a discrete variable in Netica, its conditional distribution is not represented as a step function. The reason is that each of its 10 discrete "values" is actually an interval, therefore Netica generates samples from the entire range $[0, 1]$.

In the same manner, we sample from Hours of sleep $\in [0.2, 0.3]$ and Fly duty period $\in [0.8, 0.9]$ and save the samples that Netica generates. In the simulation for vines updating, we will have to re-sample the structure, in the same conditions. For better results of the comparisons, we use the samples that we saved from Netica, in the simulation for updating with vines.

In Figure 2.6, the conditional probability tables from Netica were built using 10^4 samples. The agreement between the two methods is very poor. For example, one can notice from both curves that the combination of the two factors (not enough sleep and a long flight duty period) has an alarming effect on the crew

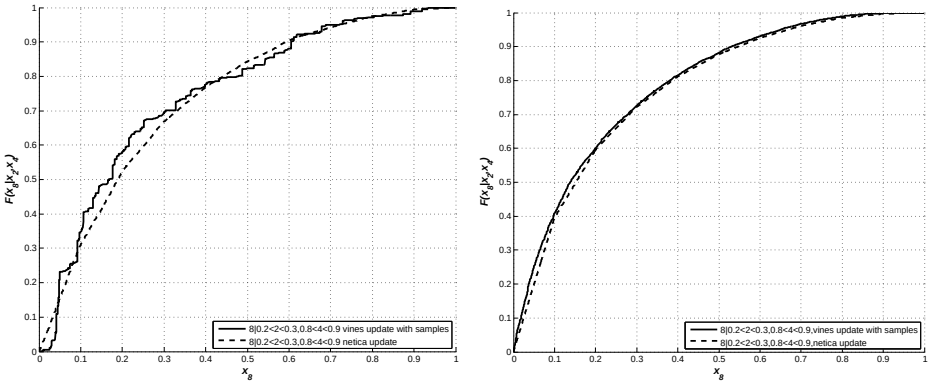


Figure 2.7: *Distribution of $X_8|X_2, X_4$. Comparison of updating results in vines and Netica using 10^4 from $8 \cdot 10^5$ samples (left) and using $8 \cdot 10^5$ samples (right).*

alertness. The difference is that in vines-updating, with probability 50 percent, alertness is less than or equal to the $15^{\underline{th}}$ percentile of its unconditional distribution[10], whereas in vines-Netica updating with probability 50 percent alertness is less than or equal to the $35^{\underline{th}}$ percentile of its unconditional distribution. This disagreement is due to the number of samples from which Netica calculates the conditional probability tables (10^4). There are 10^3 different input vectors for node 8, each requiring 10 probabilities for the distribution of 8 given the input. With 10^4 samples, we expect each of the 10^3 different inputs to occur 10 times, and we expect a distribution on 10 outcomes to be very poorly estimated with 10 samples. Moreover, updating with vines does not produce a very smooth and accurate curve, also because the simulation was performed with 10^4 samples.
In Figure 2.7 (left), the sample file imported in Netica contains $8 \cdot 10^5$ samples which allows a very good estimation of the conditional distribution of Crew alertness. Another 10^4 samples for Hours of sleep $\in [0.2, 0.3]$ and 10^4 for Fly duty period $\in [0.8, 0.9]$ are saved from Netica and used in the vines updating. The curves start to look very similar indeed, but the one corresponding to vines updating is still not smooth because of the number of samples. If we do everything with the entire sample file of $8 \cdot 10^5$ samples, the agreement between the two conditional distributions is impeccable (see Figure 2.7 (right)). This motivates the use of a very big sample file.

For a BBN with nodes that require a large number of inputs (large number of parent nodes, discretized in fairly many states) the sample files should also be very large. The big advantage is that this huge sample file needs to be done only once.

Note however that in some cases it might happen that sampling the structure,

[10] Crew alertness is a uniform variable, therefore its unconditional distribution function is the diagonal of the unit square.

even just once will cause problems, as we already mentioned in Section 2.1. We will further present a BBN structure, which at a first glance, seems very easy to deal with, in the sense that it offers a great deal of information about the dependence structure.

Example 2.2.2. *Let us consider the BBN from Figure 2.8. If the set of (conditional) rank correlations that can be elicited is either* $\{r_{21}, r_{31}, r_{42}, r_{41|2}, r_{43|21}\}$, *or* $\{r_{21}, r_{31}, r_{43}, r_{41|3}, r_{42|31}\}$, *then the BBN can be represented as one D-vine, and so the sampling procedure does not require any extra calculations. If, for some reason, these rank correlations cannot be specified, and the only ones which are available are:* $\{r_{21}, r_{31}, r_{43}, r_{42|3}, r_{41|32}\}$ *the situation worsens considerably.*
The BBN can no longer be represented as one D-vine, since the order of the variables in $\mathcal{D}^3$ *is 3, 1, 2, and in* $\mathcal{D}^4$ *is 4, 3, 2, 1.*

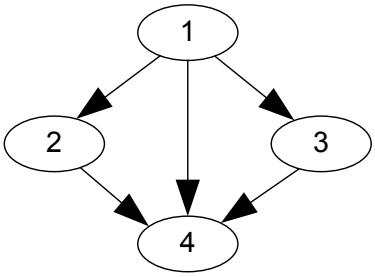


Figure 2.8: *BBN with 4 nodes and 5 arcs.*

To sample X_4, *one needs to calculate:*

$$x_4 = F^{-1}_{4|3:x_3}(F^{-1}_{4|23:F_{2|3}(x_2)}(F^{-1}_{4|123:F_{1|23}(x_1)}(u_4))).$$

The conditional distribution $F_{2|3}(x_2)$, *can be found by evaluating a double integral as in Example 2.1.1. Furthermore,* $F_{1|23}(x_1)$ *needs to be calculated. This is, in fact, the conditional distribution of* $F_{1|2}(x_1)$, *given* $F_{3|2}(x_3)$. *Even though all the information needed seems to be available, evaluating the joint distribution of these two quantile functions turns out to be very difficult. Moreover at each step of its evaluation, one should calculate the numerical value of the double integral for* $F_{3|2}(x_3)$. *This is a task that takes time and patience.*

If this kind of calculation is necessary for such a small BBN, it is very likely that more complicated ones will be involved in larger structures. The time spent to solve this sort of problems would be, by far, much longer than one can afford.

2.3 Normal Copula Vine Approach

All the troubles discussed until now are caused by the different sampling order of variables from one vine to another. To avoid these problems we advance here a new way of realizing the rank correlation specification on a regular vine using the advantages of the joint normal distribution.

Let us start with a rank correlation vine specification on the variables $X_1, \dots X_n$, with continuous, invertible distribution functions $F_1, \dots, F_n$. We adopt the following protocol:

1. Transform $X_1, \dots X_n$ to the standard normal variables $Y_1, \dots Y_n$ via the transformation $Y_i = \Phi^{-1}(F_i(X_i))$, $(\forall i)(i = 1, \dots, n)$, where Φ is the cumulative distribution function of the standard normal distribution.

2. Construct the vine for the standard normal variables $Y_1, \dots, Y_n$. Since $\Phi^{-1}(F_i(X_i))$ are strictly increasing transformations, the same (conditional) rank correlations correspond to the edges of this vine.

3. To each edge of this vine assign $\rho_{i,j|D} = 2\sin(\frac{\pi}{6} \cdot r_{i,j|D})$, where $\{i, j\}$ and D are the conditioned and conditioning sets, respectively, of the edge, and $r_{i,j|D}$ is the conditional correlation assigned to the corresponding edge from the initial vine. We now have a complete partial correlation vine specification[11] for $Y_1, \dots, Y_n$. Theorem 1.2.1 ensures that there is a unique joint normal distribution for $Y_1, \dots Y_n$ satisfying all partial correlation specifications. Moreover there is a unique correlation matrix determined by this vine (Theorem 1.2.2).

4. Compute the correlation matrix R using the recursive formula 1.2.1.

5. Sample the joint normal distribution of $Y_1, \dots, Y_n$, with correlation matrix R (Tong 1990).

6. For each sample, calculate: $(F_1^{-1}(\Phi(y_1^j)), F_2^{-1}(\Phi(y_2^j)), \dots, F_n^{-1}(\Phi(y_n^j)))$, where $((y_1^j), (y_2^j), \dots, (y_n^j))$ is the $j^{\underline{th}}$ sample from the previous step.

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

The normal copula vine method might seem very similar to the joint normal transform method presented in Ghosh and Henderson (2002); Iman and Helton (1985), but the presence of vines is crucial in avoiding the problems encountered in the latter method. In the joint normal transform approach, the rank correlation matrix must be first specified and then induced by transforming distributions to

[11] As we mentioned in Chapter 1, Section 1.2, conditional and partial correlations are equal for normal variables.

standard normals and generating a dependence structure using the linear properties of the joint normal. In absence of data, specifying a rank correlation matrix can be a very difficult task. Moreover, it is not always possible to find a product moment correlation matrix generating a given rank correlation matrix via Pearson's transformation, as showed in Chapter 4 from Kurowicka and Cooke (2006b). Using the normal copula vine approach we avoid these problems because we do not specify a rank correlation matrix, but rather a rank correlation vine, that is transformed to a partial correlation vine. All assignments of numbers between -1 and 1 to the edges of a partial correlation regular vine are consistent, in the sense that there is a joint distribution realising these partial correlations, and all correlation matrices can be obtained in this way (Bedford and Cooke 2002).

In case of a BBN which cannot be represented as one vine, we can make use of the protocol described above. Everything is calculated on the joint normal vine, hence we can reorder the variables and recompute all partial correlations needed. We expect a dramatic decrease in the computational time using this method. We note that the assumption of constant conditional rank correlations, previously a matter of convenience, is now required.

Further, we will present comparisons between the normal copula vine method and the copula-vine method together with Netica updating, using the BBN from Example 2.1.1.

The marginal distributions of X_1, X_2, X_3, X_4 are uniform on the interval $(0, 1)$. We sample the structure both with the copula-vine, and the normal copula vine approach. Hence, we produce two sample files, each containing 10^5 samples. The resulting files are imported in Netica, and conditioning is performed in both cases. Figure 2.9 presents the conditional distribution of the variable X_4 given that $X_1 \in [0.1, 0.2]$ and $X_2 \in [0.3, 0.4]$, obtained using the sample files produced with the two methods. One can notice a small disagreement between the two conditional distributions. If we think of the normal copula vine method in terms of the copula-vine method, where we made use of the normal copula, we can say that the difference between the two conditional distributions from Figure 2.9 is due to the different choice of the copula.

Another way of comparing these methods is to calculate and compare the two sample correlation matrices. The matrices presented below correspond to the sample file obtained using the copula-vine approach (left) and the sample file generated with the normal copula vine method (right):

$$\begin{pmatrix} 1 & 0.4031 & 0.7028 & 0.3746 \\ 0.4031 & 1 & 0.2843 & 0.2028 \\ 0.7028 & 0.2843 & 1 & 0.5201 \\ 0.3746 & 0.2028 & 0.5201 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0.4000 & 0.6974 & 0.3843 \\ 0.4000 & 1 & 0.2837 & 0.1985 \\ 0.6974 & 0.2837 & 1 & 0.5271 \\ 0.3843 & 0.1985 & 0.5271 & 1 \end{pmatrix}$$

Comparing the two matrices one can observe differences of order 10^{-3}, which represent a reasonable result taking into account the sampling errors.

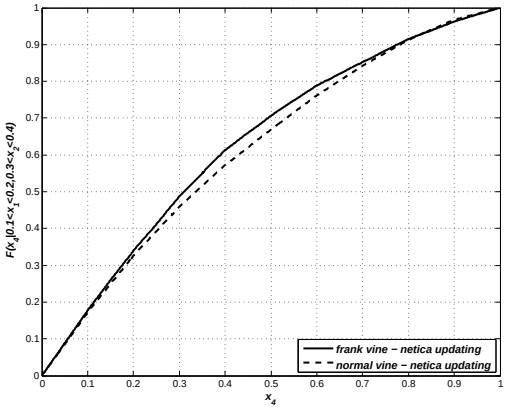


Figure 2.9: *The distribution of $X_4|X_1, X_2$. Frank's Copula Vine vs Normal Copula Vine (conditioning in Netica using 10^5 samples).*

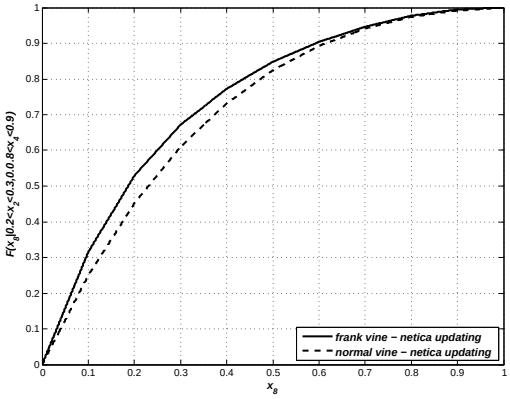


Figure 2.10: *The distribution of $X_8|X_2, X_4$. Frank's Copula Vine vs Normal Copula Vine (conditioning in Netica using $8 \cdot 10^5$ samples).*

The big advantage of the normal copula vine method is that the simulation runs for a few seconds, whereas with the previous sampling algorithm (in which a double integral needs to be numerically evaluated for each sample) the results were available in hours. Both methods were implemented in Matlab for a fair comparison of the computational times. The normal copula vine method is implemented in a software application, called UNINET. UNINET allows for quantification of non parametric continuous/discrete BBNs. The program has a friendly interface and the simulations are very fast. For details about UNINET we refer to the Appendix 7.1.

The same kind of results we find when we examine the structure from Example 2.4 (*Flight Crew Alertness*). Figure 2.10 shows the conditional distribution of the variable Crew alertness(8) given that Hour of sleep(2) is in the interval $[0.2, 0.3]$ and Fly duty period(4) $\in [0.8, 0.9]$. We can again notice that the choice of the copula produces a small discrepancy between the curves.

Comparing the two sample correlation matrices for this example we find that the maximum difference is $8 \cdot 10^{-3}$.

2.4 Analytical updating

An advantage of the normal copula vine method is that it allows for analytical conditioning. Since all the calculations are performed on a joint normal vine, any conditional distribution will also be normal. We need some notation before introducing the mean and variance of this conditional normal distribution.

Let X be a n-dimensional random vector with *multivariate normal distribution*. Let the vector μ be the expected value of X, and V be its covariance matrix. For a fixed $k < n$ consider the following partition of X, μ and V:

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, V = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb}, \end{pmatrix}$$

where $X_a = (X_1, \ldots, X_k)'$, $X_b = (X_{k+1}, \ldots, X_n)'$, $\mu_a = (\mu_1, \ldots, \mu_k)'$, $\mu_b = (\mu_{k+1}, \ldots, \mu_n)'$, $V_{ii} = var(X_{i \in \{a,b\}})$ and $V_{ab} = cov(X_a, X_b)$. The conditional variance of X_b given X_a is denoted by $\text{var}_{b|a}(X_b)$.

Proposition 2.4.1. *(Whittaker 1990) (Marginal and conditional density function): If the partitioned random vector follows the distribution:*

$$(X_a, X_b) \sim N\left[(\mu_a, \mu_b), V = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix}\right], \textit{ then:}$$

(i) the marginal distribution of X_a is normal with mean μ_a and variance V_{aa};
(ii) the conditional distribution of $(X_b|X_a)$ is normal with the mean:

$$E_{b|a}(X_b) = \mu_b + B_{b|a} \cdot (x_a - \mu_a), \textit{ where } B_{b|a} = V_{ba} V_{aa}^{-1};$$

and the variance:

$$\text{var}_{b|a}(X_b) = V_{bb|a} = V_{bb} - V_{ba}V_{aa}^{-1}V_{ab}.$$

Finding the conditional distribution of the corresponding original variable will just be a matter of transforming it back using the inverse distribution function of this variable and the standard normal distribution function.

Proposition 2.4.2. *Let (Y_1, Y_2) have a bivariate normal distribution, with standard normal marginals. Let F_1 and F_2 be two continuous, invertible distribution functions and $X_i = F_i^{-1}(\Phi(Y_i)), i = 1, 2.$, where Φ is the cumulative distribution function of the standard normal distribution. Then the conditional distribution $X_1|X_2$ can be calculated as $F_1^{-1}(\Phi(Y_1|Y_2))$.*

Proof. For $i \in \{1, 2\}$, $X_i = F_i^{-1}(\Phi(Y_i))$, therefore we can write $Y_i = \Phi^{-1}(F_i(X_i))$.

Remark: For A, B, C random variables and f a function such that $A = f(B)$, then $A|C = f(B|C)$.

We will use the above remark for X_1, X_2, Y_1, Y_2, $F_1^{-1} \circ \Phi$ and x_2 an arbitrary value of X_2 :

$$\begin{aligned} X_1|(X_2 = x_2) &= F_1^{-1}(\Phi(Y_1|(X_2 = x_2))) = F_1^{-1}(\Phi(Y_1|(F_2^{-1}(\Phi(Y_2)) = x_2))) \\ &= F_1^{-1}(\Phi(Y_1|(Y_2 = \Phi^{-1}(F_2(x_2))))) = F_1^{-1}(\Phi(Y_1|(Y_2 = y_2))), \end{aligned}$$

where we denoted $\Phi^{-1}(F_2(x_2)) = y_2$. □

Let us illustrate this result on the BBN structure from Example 2.4 (*Flight Crew Alertness*). Figure 2.11 presents a comparison between updating in the normal copula vine method and the copula-vine method. The updating is performed both in Netica and analytically.
As one would expect, the pairs of curves corresponding to the two methods (copula-vine and normal copula vine) follow exactly the same patterns regardless of the way we perform conditioning. The distance between the pairs of curves is caused by the different choice of the copula.

We will now consider (for the same example) the univariate distributions to be standard normals instead of uniforms on $(0, 1)$. The same kind of comparisons as before are performed. In doing so, a new model should be pre-prepared in Netica. The differences between the new model and the one presented in Figure 2.5 are the range of the variables and the discretization intervals. We will keep the same number of intervals for the discrete version of each variable, only they will not be equally sized anymore. The variables are discretized with respect to the quantiles of their distributions.

We conditionalize on Hours of sleep between its 0.2 and 0.3 quantiles and Fly

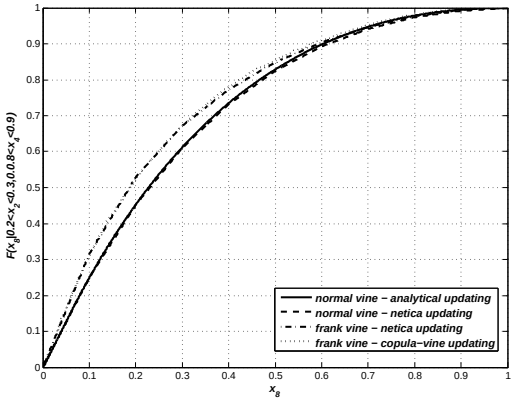


Figure 2.11: *The distribution of $X_8|X_2, X_4$. Comparison of updating results in Frank's Copula Vine (using Netica and the copula-vine updating) vs updating in Normal Copula Vine (using Netica and analytically). All univariate distributions are uniforms on* $(0, 1)$.

duty period between its 0.8 and 0.9 quantiles[12]. The conditional distribution of the Crew alertness, obtained with the methods previously discussed, is presented in Figure 2.12(left).

The curves nicely agree everywhere, except for the first interval of the discretization, where the results given by Netica updating, in both methods, are completely different from the results of the analytical updating[13]. As already stated, the discretization of the variables was made according to their quantiles, hence the first interval and the last one (for each variable) are much wider than the rest of the intervals. This can be noticed in Figure 2.12(right), which shows the *Flight Crew Alertness* structure in Netica, after we updated the model. A sample file of $8 \cdot 10^5$, obtained with the normal copula vine method, was imported in Netica in order to create the conditional probability tables. For the variable Crew alertness, the first and the last discretization intervals are approximately 12 times wider than the rest of the intervals from its discretization (see Figure 2.12(right)). In order to plot the conditional distribution of Crew alertness given by Netica, one will need to generate samples from it. Netica simply samples uniformly from each discretization interval, taking into account its probability. The information that most of the samples from the first interval should be concentrated in its right hand side, is not included. Therefore, the first part of each of the curves given by Netica does not resemble reality. The same kind of discrepancy would happen in

[12]In the previous comparisons(for uniform marginals) the conditioning was Hours of sleep $\in [0.2, 0.3]$ and Fly duty period $\in [0.8, 0.9]$.

[13]The word "analytical" is appropriate only for the normal copula vine method. For the copula-vine method, updating is performed via re-sampling the structure.

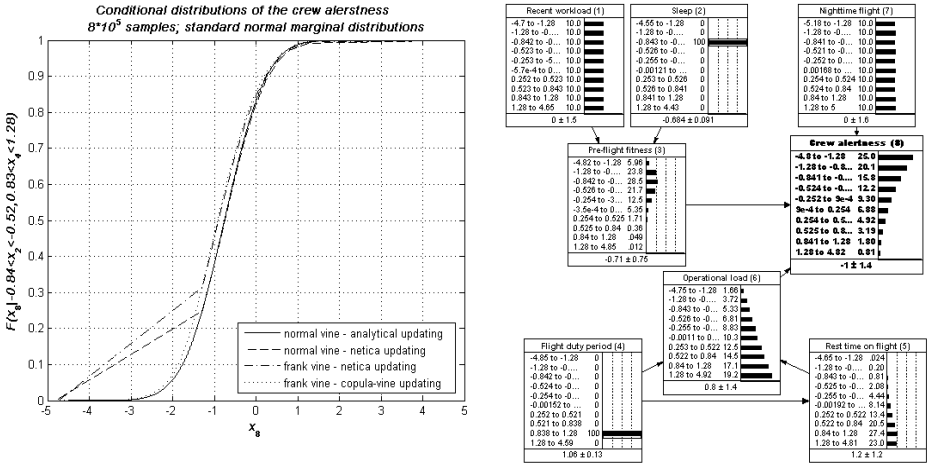


Figure 2.12: *The distribution of $X_8|X_2, X_4$. Comparison of updating results in Frank's Copula Vine (using Netica and the copula-vine updating) vs updating in Normal Copula Vine (using Netica and analytically)(left). The Flight Crew Alertness model in Netica(right). All univariate distributions are standard normals.*

the last interval if its probability were larger.

We will further consider another updating of the same model. We conditionalize on Hours of sleep between its 0.3 and 0.4 quantiles, Operational load between its 0.3 and 0.4, and Nighttime flight between its 0.4 and 0.5 quantiles. Figure 2.13(right) shows the structure in Netica, after updating. Looking at the conditional probability of Crew alertness, one can notice that the first and the last discretization intervals have very small probabilities. In these conditions, we expect the curves for the conditional distribution of the Crew alertness, obtained with the four methods, to be very similar on the entire domain. As Figure 2.13(left) shows, there is perfect agreement between the methods.

Although in most cases Netica updating has a reasonable outcome, in some particular ones, its results are not to be trusted. Therefore the opportunity to perform analytical updating is a big advantage of the normal copula vine method. Moreover, if the BBN contains some nodes, each with a very large number of inputs, Netica will not have enough memory to store their respective CPT's.

The analytical updating using the normal copula vine method can be also performed in UniNet. This offers the major advantage to work with the continuous variables, rather than with their discretized version. Figure 2.14 shows the histograms of the standard normal variables from the *Flight Crew Alertness* model. The means and standard deviations are displayed under the histograms.

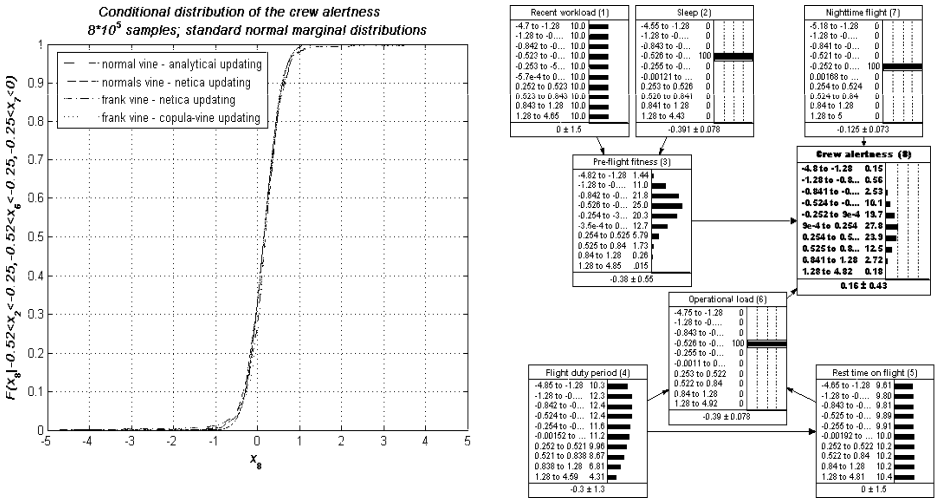


Figure 2.13: *The distribution of $X_8|X_2, X_6, X_7$. Comparison of updating results in Frank's Copula Vine (using Netica and the copula-vine updating) vs updating in Normal Copula Vine (using Netica and analytically)(left). The Flight Crew Alertness model in Netica(right). All univariate distributions are standard normals.*

Figure 2.15 displays the result of the conditionalisation on single values of the variables Hours of sleep, Operational load, and Nighttime flight. The conditionalisation from Figure 2.15a is similar to the one showed in Figure 2.13, whereas in Figure 2.15b we conditionalise on more extreme values of the same variables. The grey distributions in the background are the unconditional marginal distributions, provided for comparison. The conditional means and standard deviations are also displayed.

The theory presented here can be extended to include ordinal discrete random variables; that is variables which can be written as monotone transforms of uniform variables, perhaps taking finitely many values. The dependence structure must be defined with respect to the uniform variates. The case of dichotomous variables (and generally variables with few states), and the rank correlations between two such variables is investigated in the next chapter.

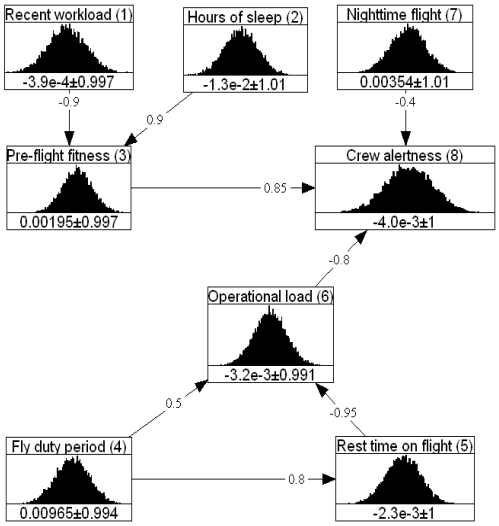
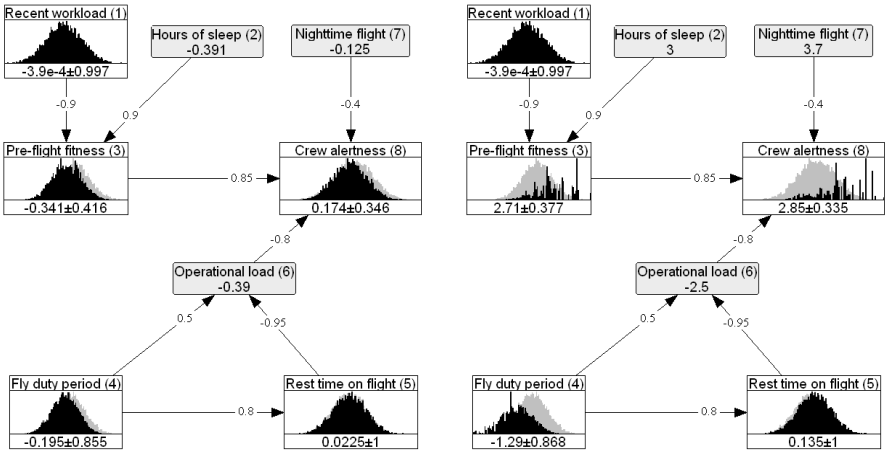


Figure 2.14: *Flight crew alertness model with histograms in* UniNet.



(a) $X_2 = -0.391$, $X_6 = -0.39$ and $X_7 = -0.125$.

(b) $X_2 = 3$, $X_6 = -2.5$ and $X_7 = 3.7$.

Figure 2.15: *The conditional distribution given different values of* X_2, X_6 *and* X_7.

Chapter 3

Spearman's Rank Correlation for Ordinal Discrete Random Variables[1]

In order to extend the approach to non-parametric continuous BBNs, such that it includes ordinal discrete random variables, we need to study the concept of rank correlation between two such variables.

A population version of Spearman's rank correlation has been defined in the case of continuous variables. Our interest is in the discrete case. We start this chapter with an overview of existing rank correlation measures between two discrete random variables. Section 3.2 introduces a number of definitions and preliminary results necessary further in this chapter. In Section 3.3 we propose a generalisation of the population version of Spearman's rank for the case of ordinal discrete random variables.

Discrete univariate distributions can be obtained as monotone transforms of uniform variables. A class of discrete bivariate distributions can be constructed by specifying the marginal distributions and a copula. In contrast with the continuous case, the rank correlation coefficient of the discrete variables depends on not only the copula, but also the marginal distributions. In Section 3.4 we give the analytical description of the dependence between the rank correlation of two discrete variables (with given marginals) and the rank correlation of their underlying uniforms. This relation is needed in our BBN modeling approach, where the dependence structure is defined with respect to the uniform variates. We study this relationship for different copulae and different marginal distributions. For certain choices of marginals, the correlation of the discrete variables does not attain the whole range of dependence.

[1]This chapter is based on the paper Hanea et al. (2007), "The Population Version of Spearman's Rank Correlation Coefficient in the Case of Ordinal Discrete Random Variables", published in the Proceedings of the Third Brazilian Conference on Statistical Modelling in Insurance and Finance.

3.1 Context

In many practical problems one needs to quantify the dependence structure among variables. Various dependence measures have been proposed and studied. Rank-based dependence criteria, such as Spearman's rank correlation r, and Kendall's τ, are usually used to measure dependence in bivariate responses. However, in most cases, the variables involved in the analysis are continuous. The population versions of the rank correlations are formulated in the continuous case only. We are interested in the discrete case.

The matter of describing dependence between two discrete random variables in terms of rank correlation has been receiving much attention lately, see for example Vandenhende and Lambert (2003), Mesfioui and Tajar (2005), Denuit and Lambert (2005), and Neslehova (2007). The approaches taken in Mesfioui and Tajar (2005), Denuit and Lambert (2005) and Neslehova (2007) are very similar, and they are based on a continuousation principle, i.e. a transformation of an arbitrary random variable to a continuous variable. The population version of Spearman's rank correlation derived in Neslehova (2007) coincides with the one derived in this chapter. In addition Neslehova (2007) proves that the sample version of the derived formula is precisely the sample version of Spearman's rank correlation with the classical tie correction. Bearing in mind that Spearman's rank correlation appeared as a measure for discrete random variables, we approach the problem in a somewhat reversed way.

A correction for Kendall's τ was introduced in Goodman and Kruskal (1954), in order to adjust this measure for ties. This was further studied and detailed in Vandenhende and Lambert (2000). We use similar techniques as in Vandenhende and Lambert (2000) to study Spearman's rank correlation coefficient for discrete variables. We calculate a correction for the population version of Spearman's r, starting from its sample version, when ties are present. The resulting expression is proportional with the difference between the probabilities of concordance and discordance. The proportionality factor is a function of marginal distributions. With continuous responses, this factor equals 3, and the formula reduces to the standard definition of Spearman's rank correlation r.

It is well known that the rank correlation in the discrete case does not necessarily attain the whole range of dependence $[-1, 1]$. We propose a way to deal with this problem, so that the same ± 1 limits are always reached under complete dependence. This way is similar with the approach adopted in Vandenhende and Lambert (2003), in the sense that it uses the Fréchet bounds of a copula in order to determine a scaling factor for the dependence measure. In the case of binary variables, the alternative form of Spearman's rank correlation proposed in Vandenhende and Lambert (2003), and the normalised correction for the population version of Spearman's r proposed here are identical.

3.2 Definitions & Concepts

3.2.1 The population version of Spearman's r for continuous variables

Consider a population distributed according to two variables 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):

$$r = 3 \cdot (P_c - P_d). \tag{3.2.1}$$

The above definition of Spearman's rank correlation is appropriate only for populations for which the probabilities of $X_1 = X_2$ or $Y_1 = Y_2$ are zero. Such populations are mainly infinite populations with both X and Y distributed continuously (Hoffding 1947). We will further present a simple example in which the rank correlation given by formula (3.2.1) does not describe the dependence structure as expected.

Remark 3.2.1. *Let us consider two binary variables X and Y with the following joint distribution:*

$X \setminus Y$	1	2	
1	0.5	0	0.5
2	0	0.5	0.5
	0.5	0.5	

Table 3.1: *Joint distribution of two completely plosively correlated binary variables.*

The above structure suggests that X and Y are completely positively correlated, yet calculating their rank correlation using formula (3.2.1), we obtain $r = 0.75$.

In order to formulate a population version of Spearman's r for discrete variables, one will have to correct for the probabilities of ties, since in the discrete case $P(X_1 = X_2) > 0$ and $P(Y_1 = Y_2) > 0$. This correction is similar to the correction for ties for the sample version of the rank correlation.

3.2.2 The sample version of Spearman's r in the presence of ties

The early references for Spearman's rank correlation coefficient are Spearman (1904) and Spearman (1906). Spearman proposed a method to determine correlation, based on replacing measurements with their ranks. We will create the necessary mathematical context to introduce this method.

Let us consider N samples of the random vector (X, Y): $(x_1, y_1), \ldots, (x_N, y_N)$. Suppose the samples for both variables are ranked[2]. For each value x_k we define the rank s_k, and for each value y_k we define the rank r_k. The rank correlation coefficient proposed in (Spearman 1906) may be regarded as the product moment correlation between the ranks of two variables. We will denote it r_s^N to indicate that is a sample version for N samples. The variance of a set of values which are the first N integers is calculated as $\frac{1}{N}\sum_{i=1}^{N} i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} i\right)^2 = \frac{N^2-1}{12}$. Therefore r_s^N is given by:

$$r_s^N = \frac{12}{N^3 - N}\sum_{i=1}^{N}\left(s_i - \frac{N+1}{2}\right)\left(r_i - \frac{N+1}{2}\right).$$

This formula can be reduced to a more familiar form, if we denote D^2 the sum of squared differences between the ranks of each pair, i.e. $D^2 = \sum_{k=1}^{N}(s_k - r_k)^2$:

$$r_s^N = 1 - \frac{6D^2}{N^3 - N}. \tag{3.2.2}$$

Let us now assume that there are tied values in the sequences $(x_i)_{i=1,\ldots,N}$ and $(y_i)_{i=1,\ldots,N}$. Hence there are sets of tied ranks in the rankings of each variable. Following the analogy with the product moment correlation, "Student" ("Student" 1921) showed that the effect of tied ranks is to modify the variance of the ranking. He proposed a correction for r_s^N in the case of tied ranks. We shall further present this correction, with more emphasis on the notation and its final form, than on how it was derived.

Let us divide the sequence $(x_i)_{i=1,\ldots,N}$ into $m < N$ blocks of identical values[3]. Let u_i denote the number of values in the i^{th} block. In the same way we consider $n < N$ blocks of identical values in the sequence $(y_i)_{i=1,\ldots,N}$, and we denote with v_j the number of values in the j^{th} block. It follows that in the rankings of each

[2]When objects are arranged in order according to some quality which they all posses to a varying degree, they are said to be *ranked* with respect to that quality (Kendall and Gibbons 1990).

[3]We will consider single values as blocks containing one value.

variable there are sets of u_i and v_j tied ranks, respectively. We define:

$$U = \frac{1}{12}\sum_{i=1}^{m}(u_i^3 - u_i); \quad V = \frac{1}{12}\sum_{j=1}^{n}(v_j^3 - v_j).$$

The *midrank method* is used for assigning ranks to tied values[4]. This method is to average the ranks which they would posses if they were not tied. For example, if the fifth and the sixth members (in an ordered sequence) are tied, each is assigned the rank $5\frac{1}{2}$. In general, if ties occur for the i^{th} to the k^{th} inclusive members, the midrank is $\frac{i+k}{2}$. There are m distinct ranks for X, and n distinct ranks for Y. For the sake of coherence, it will be convenient to reindex the sequences of distinct ranks in the following way: $(s_i)_{i=1,\dots,m}$ for X and $(r_j)_{j=1,\dots,n}$ for Y. The ranks can be calculated as follows:

$$\begin{aligned} s_i &= \sum_{k=1}^{i-1} u_k + \frac{u_i+1}{2} = \frac{N+1}{2} + \frac{1}{2}\sum_{k=1}^{i-1} u_k - \frac{1}{2}\sum_{k=i+1}^{m} u_k; \\ r_j &= \sum_{j=1}^{l-1} v_l + \frac{v_j+1}{2} = \frac{N+1}{2} + \frac{1}{2}\sum_{l=1}^{j-1} v_l - \frac{1}{2}\sum_{l=j+1}^{n} v_l. \end{aligned} \tag{3.2.3}$$

Let us denote with S_{ij} the number of occurrences of the pair of ranks (s_i, r_j). Then:

$$D^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} S_{ij}(s_i - r_j)^2. \tag{3.2.4}$$

The information given by the sample $(x_1, y_1), \dots, (x_N, y_N)$ together with its ranks can be written in a more compact way, as shown in Table 3.2.

$X \setminus Y$	r_1	r_2	...	r_n	
s_1	S_{11}	S_{12}	...	S_{1n}	u_1
s_2	S_{21}	S_{22}	...	S_{2n}	u_2
...	...	...	...	...	...
s_m	S_{m1}	S_{m2}	...	S_{mn}	u_m
	v_1	v_2	...	v_n	N

Table 3.2: *Sample distribution of ranks for* (X, Y)

[4]An alternative to the midrank method is the *bracket-rank method*, to which "Student" ("Student" 1921) referred as a suggestion of Karl Pearson. In this method the tied values are all ranked as if they were the highest member of the tie. The disadvantages of this method are discussed in (Kendal 1945).

Remark 3.2.2. *Notice that* $u_i = \sum_{j=1}^{n} S_{ij}$ *and* $v_j = \sum_{i=1}^{m} S_{ij}$. *Moreover* $\sum_{i=1}^{m} u_i = \sum_{j=1}^{n} v_j = N$.

We will denote with r_{st}^{N} [5] the rank correlation for samples, when ties are present. This is defined as follows ("Student" 1921):

$$r_{st}^{N}(X,Y) = \frac{\frac{1}{6}(N^3 - N) - D^2 - U - V}{\sqrt{\left(\frac{1}{6}(N^3 - N) - 2U\right)\left(\frac{1}{6}(N^3 - N) - 2V\right)}}. \tag{3.2.5}$$

In order to formulate the population version of Spearman's rank correlation coefficient of two discrete variables, we start from the sample version of the rank correlation when ties are present, given in formula (3.2.5).

3.3 The population version of Spearman's r for ordinal discrete variables

Let us now 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. 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.

$X_1 \setminus Y_1$	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}	

Table 3.3: *Joint distribution of* (X_1, Y_1) *(left); Joint distribution of* (X_2, Y_2) *(right)*

In Table 3.3, 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}, $j = 1, ..., n$. Each q_{ij} can be rewritten as $q_{ij} = p_{i+}p_{+j}$, for all $i = 1, ..., m$ and $j = 1, ..., n$. Using this terminology we calculate:

$$P_c - P_d = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}\left(\sum_{k\neq i}\sum_{l\neq j} sign\left((k-i)(l-j)\right)q_{kl}\right)\right). \tag{3.3.1}$$

[5] The extra index t indicates the presence of ties.

Spearman's rank correlation coefficient of two discrete variables can be calculated using the following theorem:

Theorem 3.3.1. *Consider a population distributed according to two variables X and Y. Let (X_1, Y_1), as above, be a member of this population. Let (X_2, Y_2) satisfy the conditions described above. Then sample version of Spearman's rank correlation coefficient $r_{st}^N(X,Y)$, of X and Y, given by (3.2.5) converges to:*

$$\bar{r} = \frac{3\left(P_c - P_d\right)}{\sqrt{\left(1 - \sum_{i=1}^{m} p_{i+}^3\right) \cdot \left(1 - \sum_{j=1}^{n} p_{+j}^3\right)}},$$

when $N \to \infty$. p_{i+} and p_{+j} are given by Table 3.3.

Proof: Let $P_c - P_d$ be given by formula (3.3.1). We start from the sample version of the rank correlation when ties are present given by formula (3.2.5). Divide by N the entries of Table 3.2. When $N \to \infty$ the result of this division will approximate the joint distribution of a random vector (X, Y), given as in Table 3.3 (left). Hence:

$$\frac{S_{ij}}{N} \to p_{ij}; \quad \frac{u_i}{N} \to p_{i+}; \quad \textit{and} \quad \frac{v_j}{N} \to p_{+j} \quad (\forall i, j)\,(i = 1, \ldots, m, j = 1, \ldots, n)\,. \tag{3.3.2}$$

Let us start by rewriting formula (3.2.5) in a more convenient way:

$$r_{st}^N(X,Y) = \frac{\frac{1}{6}N^3 - \frac{1}{6}N - D^2 - \frac{1}{12}\sum_{i=1}^{m} u_i^3 + \frac{1}{12}\sum_{i=1}^{m} u_i - \frac{1}{12}\sum_{j=1}^{n} v_j^3 + \frac{1}{12}\sum_{j=1}^{n} v_j}{\sqrt{\left(\frac{1}{6}N^3 - \frac{1}{6}N - \frac{1}{6}\sum_{i=1}^{m} u_i^3 + \frac{1}{6}\sum_{i=1}^{m} u_i\right)\left(\frac{1}{6}N^3 - \frac{1}{6}N - \frac{1}{6}\sum_{j=1}^{n} v_j^3 + \frac{1}{6}\sum_{j=1}^{n} v_j\right)}}.$$

Using Remark 3.2.2 and dividing by N^3 we obtain:

$$r_{st}^N(X,Y) = \frac{\frac{1}{6} - \frac{1}{12}\sum_{i=1}^{m}\left(\frac{u_i}{N}\right)^3 - \frac{1}{12}\sum_{j=1}^{n}\left(\frac{v_j}{N}\right)^3 - \frac{D^2}{N^3}}{\frac{1}{6}\sqrt{\left(1 - \sum_{i=1}^{m}\left(\frac{u_i}{N}\right)^3\right)\left(1 - \sum_{j=1}^{n}\left(\frac{v_j}{N}\right)^3\right)}}. \tag{3.3.3}$$

Let us denote:

$$\bar{U} = \frac{1}{12}\sum_{i=1}^{m}\left(\left(\frac{u_i}{N}\right)^3 - \frac{u_i}{N}\right); \quad \bar{V} = \frac{1}{12}\sum_{j=1}^{n}\left(\left(\frac{v_j}{N}\right)^3 - \frac{v_j}{N}\right). \tag{3.3.4}$$

Using the above notations, we can rewrite relation (3.3.3) as follows:

$$r_{st}^N = \frac{-\bar{U} - \bar{V} - \frac{D^2}{N^3}}{\sqrt{(-2\bar{U})(-2\bar{V})}}. \tag{3.3.5}$$

From equation (3.2.4) we have:

$$\frac{D^2}{N^3} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{S_{ij}}{N}\left(\frac{s_i}{N} - \frac{r_j}{N}\right)^2. \tag{3.3.6}$$

If we further use the result from (3.3.2) in formulae (3.3.4) and (3.3.6), we can rewrite:

$$\bar{U} = \frac{1}{12}\sum_{i=1}^{m}\left(p_{i_+}^3 - p_{i_+}\right); \quad \bar{V} = \frac{1}{12}\sum_{j=1}^{n}\left(p_{+j}^3 - p_{+j}\right); \tag{3.3.7}$$

and

$$\frac{D^2}{N^3} = \frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\left(\sum_{k=1}^{i-1} p_{k_+} - \sum_{k=i+1}^{m} p_{k_+}\right) + \left(\sum_{l=j+1}^{n} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)\right)^2. \tag{3.3.8}$$

Hence we can express r_{st}^N only in terms of p_{ij}, p_{i_+} and p_{+j}. The resultant expression for the rank correlation will be denoted with $\bar{r}$. Without loss of generality we will further consider $n = m$, and proceed with the calculations:

$$\begin{aligned} -\bar{U} &= \frac{1}{12}\left(\left(p_{1_+} + p_{2_+} + \ldots + p_{m_+}\right)^3 - \left(p_{1_+}^3 + p_{2_+}^3 + \ldots + p_{m_+}^3\right)\right) \qquad (3.3.9)\\ &= \frac{1}{4}\left(\sum_{i=1}^{m}\sum_{i\neq j} p_{i_+}^2 p_{j_+} + 2\sum_{k>j>i} p_{i_+}p_{j_+}p_{k_+}\right). \end{aligned}$$

In the same manner we obtain the following for $-\bar{V}$:

$$-\bar{V} = \frac{1}{4}\left(\sum_{i=1}^{m}\sum_{i\neq j} p_{+i}^2 p_{+j} + 2\sum_{k>j>i} p_{+i}p_{+j}p_{+k}\right). \tag{3.3.10}$$

From formula (3.3.8) we have:

$$\begin{aligned} \frac{D^2}{N^3} &= \frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k=1}^{i-1} p_{k_+} - \sum_{k=i+1}^{m} p_{k_+}\right)^2 + \frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)^2\\ &+ \frac{1}{2}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k=1}^{i-1} p_{k_+} - \sum_{k=i+1}^{m} p_{k_+}\right)\left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right). \qquad (3.3.11) \end{aligned}$$

One can recalculate the first term of the above sum and, using formula (3.3.9), show the following:

$$\begin{aligned}\frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k=1}^{i-1} p_{k+} - \sum_{k=i+1}^{m} p_{k+}\right)^2 &= \frac{1}{4}\sum_{k=1}^{m} p_{k+}^2 \cdot \sum_{i=1}^{m}\sum_{k\neq j} p_{ij} - \frac{1}{2}\sum_{k<j<l} p_{k+}p_{l+}p_{j+} \\ &+ \frac{1}{2}\sum_{k<l<j} p_{k+}p_{l+}p_{j+} + \frac{1}{2}\sum_{j<l<k} p_{k+}p_{l+}p_{j+} \\ &= \frac{1}{4}\left(\sum_{k=1}^{m}\sum_{k\neq j} p_{k+}^2 p_{j+} + 2\sum_{k<l<j} p_{k+}p_{l+}p_{j+}\right) \\ &= -\bar{U}. \end{aligned} \tag{3.3.12}$$

Recalculating the second sum and using (3.3.10) we obtain:

$$\frac{1}{4}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{l=j+1}^{m} p_{+l} - \sum_{l=1}^{j-1} p_{+l}\right)^2 = \frac{1}{4}\left(\sum_{l=1}^{m}\sum_{l\neq j} p_{+l}^2 p_{+j} + 2\sum_{l<k<j} p_{+k}p_{+l}p_{+j}\right) = -\bar{V}. \tag{3.3.13}$$

The last term of (3.3.11) can be also rewritten as:

$$-\frac{1}{2}\sum_{i,j=1}^{m} p_{ij}\left(\sum_{k\neq i}\sum_{l\neq j} sign\left((k-i)(l-j)\right) q_{kl}\right). \tag{3.3.14}$$

Using relations (3.3.12), (3.3.13) and (3.3.14) in equation (3.3.11) we can write:

$$\frac{D^2}{N^3} = -\bar{U} - \bar{V} - \frac{1}{2}(P_c - P_d).$$

Therefore formula (3.3.5) becomes:

$$\bar{r}(X,Y) = \frac{\frac{1}{2}(P_c - P_d)}{\sqrt{(-2\bar{U})(-2\bar{V})}}.$$

If we further use relation (3.3.7) we obtained the desired expression. □

We shall call $\bar{r}$ *the population version* of Spearman's rank correlation coefficient of X and Y, whose sample version is given by (3.2.5).

One can express $\sum_{i=1}^{m} p_{i+}^3$ as $P(X_1 = X_2 = X_3)$ where X_1, X_2 and X_3 are independent with the same distribution. For continuous variables $P(X_1 = X_2 = X_3) = 0$ and the denominator of the formula given in Theorem 3.3.1 is 1. Hence, in this case, $\bar{r}$ is equivalent to r for continuous variables. It is worth mentioning that the population version of Spearman's r was initially defined using 3 pairs of independent copies of (X, Y).

Remark 3.3.1. *Spearman's rank correlation of two continuous variables X and Y, with cumulative distribution functions F_X and F_Y, is the product moment correlation of $F_X(X)$ and $F_Y(Y)$, i.e. $\rho(F_X(X), F_Y(Y))$. If X and Y are discrete variables, calculating the product moment correlation of $F_X(X)$ and $F_Y(Y)$ will yield a different result than $\bar{r}$. This difference comes from the different method of dealing with tied values. In the derivation of $\bar{r}$ we used the midrank method, whereas applying the cumulative distribution function is equivalent with using the bracket-rank method mentioned in Section 3.2.2. The disadvantages of this method are discussed in (Kendal 1945). Nevertheless, we will present here an example in which the dependence structure is described better by $\bar{r}$ than by $\rho(F_X(X), F_Y(Y))$.*

Let us consider two binary variables X and Y distributed as in Table 3.4.

$X \setminus Y$	1	2	3	
1	0	0	0.2	0.2
2	0.2	0.1	0	0.3
3	0.5	0	0	0.5
	0.7	0.1	0.2	

Table 3.4: *Joint distribution of two strongly negatively correlated discrete random variables.*

This structure suggests that X and Y are strongly negatively correlated. This is confirmed when calculating $\bar{r}$ = -0.813. Yet calculating $\rho(F_X(X), F_Y(Y))$ we obtain only -0.173.

3.4 Dependence models using copulae

Univariate discrete distributions can be obtained as monotone transforms of uniform variables. A class of bivariate discrete distributions can be constructed by specifying the marginal distributions and a copula.

Each term p_{ij} from Table 3.3 (left) can be written in terms of the chosen copula, as follows[6]:

$$\begin{aligned} p_{ij} &= C\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) + C\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right) \\ &- C\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - C\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right). \end{aligned} \tag{3.4.1}$$

[6]The class of discrete distributions that we obtain in this way will depend on the choice of the copula and its properties.

One-parameter copulae can be parameterised by their rank correlation r, so we will use the notation C_r instead of C. Further, we will establish the relation between the rank correlation of the discrete variables and the rank correlation of the underlying uniforms.

Theorem 3.4.1. *Let C_r be a copula and (X,Y) a random vector distributed as in Table 3.3 (left), where each p_{ij} is given by formula (3.4.1) . Then the rank correlation of X and Y is denoted $\bar{r}_C$ and it can be calculated as $\bar{r}$ in Theorem 3.3.1, where:*

$$P_c - P_d = \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} p_{i+}p_{+j}\left(\widetilde{C}^r_{ij} - 1\right) \quad \textit{and} \tag{3.4.2}$$

$$\begin{aligned}\widetilde{C}^r_{ij} &= C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) + C_r\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j} p_{+l}\right)\\ &+ C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right) + C_r\left(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\right).\end{aligned}$$

Moreover, if C_r is a positively ordered copula[7], then $\bar{r}_C$ is an increasing function of the rank correlation of the underlying uniforms.

Proof: For simplicity, all calculations will be done for the case $n = m$. In order to prove the expression of $P_c - P_d$ from the theorem, we first acquire an intermediate result:

$$\begin{aligned}P_c - P_d &= \sum_{i,j=1}^{m-1} p_{ij}\left(p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+}\right)\left(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}\right)\\ &- \sum_{i,j=1}^{m-1} p_{i+}p_{+j}.\end{aligned} \tag{3.4.3}$$

To do so, we start from equation (3.3.1) and rewrite the double sum in terms of p_{ij}, p_{i+}, p_{+j} with $i, j = 1, .., m-1$. Collecting alike terms and performing a number of calculations we obtain equation (3.4.3). Now we can use the expression for p_{ij} from (3.4.1) to rewrite the first part[8] of equation (3.4.3). After algebraic manipulations of the terms we obtain[9]:

$$P_c - P_d = \sum_{i,j=1}^{m-1}\left(p_{i+} + p_{(i+1)+}\right)\left(p_{+j} + p_{+(j+1)}\right) C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j}. \tag{3.4.4}$$

[7]The definition of positively ordered copulae can be found in Chapter 1.
[8]Notice that the second part of the equation is in the proper form.
[9]All calculations can be found in Appendix 7.2.

The expression in (3.4.2) is obtained by simply rearranging the terms from the above equation.

Let C_r be a positively ordered copula. Then, from the above expression $\bar{r}_C$ is a linear combination, with positive coefficients, of positively ordered copulae. Hence the rank correlation of two discrete variables is an increasing function of the rank correlation of the underlying uniforms. □

The expression of $P_c - P_d$ given by formula (3.4.2) is also derived in Conti (1993), Kowalczyk and Niewiadomska-Bugaj (2001), and Mesfioui and Tajar (2005), in 3 different ways. The approach used here is completely different from all of them, and allows for the use of this formula in connecting the rank correlation of two discrete variables with that of their underlying uniforms.

If we look at the limiting case, when $m, n \to \infty$, and each $p_{ij} \to f(x_i, y_j)dx_i dy_j$, the expression for $P_c - P_d$ given in Theorem 3.4.1 is equivalent to:

$$\iint 4 f_X(x) f_Y(y) C_r(F_X(x), F_Y(y)) d_x d_y - 1. \tag{3.4.5}$$

where f_X, f_Y are the marginal densities of X and Y, respectively, and F_X and F_Y are the marginal distributions of X and Y, respectively. If we denote $F_X(X) = U$ and $F_Y(Y) = V$, then expression (3.4.5) becomes $4 \iint C_r(u, v) d_u d_v - 1$, which is equal to $P_c - P_d$ for continuous variables (Nelsen 1999).

Remark 3.4.1. *If we consider only $m \to \infty$ and each $p_{i+} \to f_X(x_i)dx_i$, Theorem 3.4.1 allows us to calculate the rank correlation between a continuous and a discrete random variable as follows:*

$$\bar{r}_C = \frac{3\,(P_c - P_d)}{\sqrt{1 - \sum_{j=1}^{n} p_{+j}^3}},$$

where

$$P_c - P_d = 2 \int f_X(x) \sum_{j=1}^{n-1} p_{+j} \left(C_r \left(F_X(x), \sum_{l=1}^{j} p_{+l} \right) + C_r \left(F_X(x), \sum_{l=1}^{j-1} p_{+l} \right) - 1 \right) d_x.$$

Note that any copula can be used in expression (3.4.2) from Theorem 3.4.1. If the independence 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.

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. It is worth pointing out that the copulae used in our analysis allow a full range of positive and negative dependence, have reflection symmetry[10] (Joe 1997), and that zero correlation entails the independence copula.

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 positively 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[11] to r.

When marginal distributions are not uniform, the relationship changes. Figure 3.1 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[12]: $p_{1+} = p_{+1} = 0.01$; $p_{2+} = p_{+2} = 0.98$ and $p_{3+} = p_{+3} = 0.01$. We use Frank's copula to obtain Figure 3.1a, and the Normal copula in Figure 3.1b.

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 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 the origin 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 3.2a.

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 3.2b.

Let us now consider the same marginal distributions as in Figures 3.2a and 3.2b, and use a copula which is not positively ordered. We choose Mardia's copula, which is neither positively, nor negatively ordered (e.g. Nelsen 1999).

Since Mardia's copula is not positively ordered, $\bar{r}_C$ is not an increasing function of r anymore (see Figure 3.3).

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 3.4a) that both positive and negative dependencies are bounded.

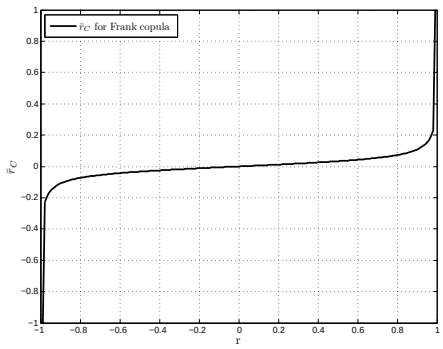
Using the Fréchet bounds for C_r in expression (3.4.4):

$$\max\left(0, \sum_{k=1}^{i} p_{k+} + \sum_{l=1}^{j} p_{+l} - 1\right) \leq C_r\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right) \leq \min\left(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\right),$$

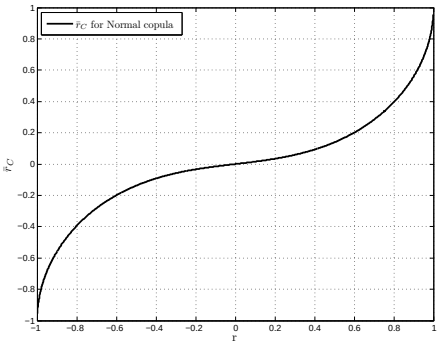
[10] The *reflection symmetry* property is called *radial symmetry* in Nelsen (1999).

[11] 10 states for each variable will suffice to obtain differences of order 10^{-3} between $\bar{r}_C$ and r.

[12] We use the notation from Table 3.3 to describe the marginal distributions of X and Y.

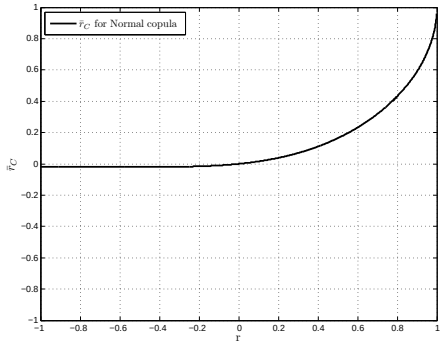


(a) $p_{1+} = p_{+1} = p_{3+} = p_{+3} = 0.01, p_{2+} = p_{+2} = 0.98$. The joint distribution is constructed using Frank's copula.

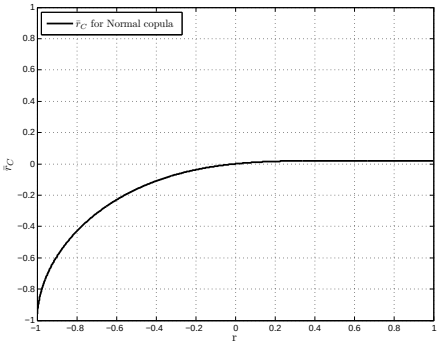


(b) $p_{1+} = p_{+1} = p_{3+} = p_{+3} = 0.01, p_{2+} = p_{+2} = 0.98$. The joint distribution is constructed using the Normal copula.

Figure 3.1: *The relationship between the parameter r, of a chosen copula, and $\bar{r}_C$, for discrete variables with equal and symmetric marginal distributions.*

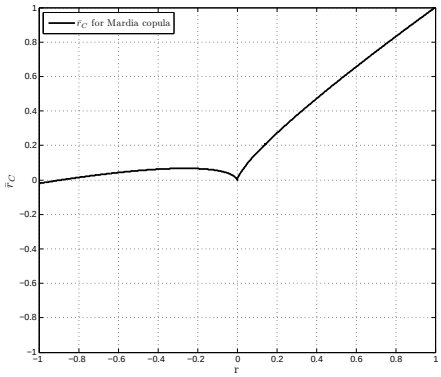


(a) $p_{1+} = p_{+1} = p_{2+} = p_{+2} = 0.01, p_{3+} = p_{+3} = 0.98$.

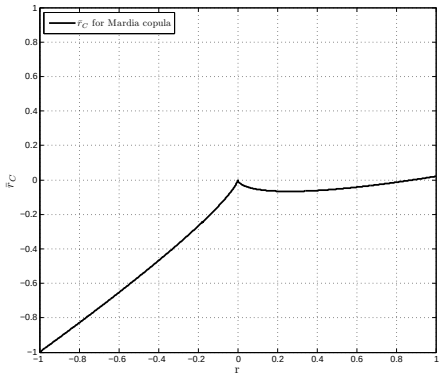


(b) $p_{2+} = p_{3+} = p_{+1} = p_{+2} = 0.01, p_{1+} = p_{+3} = 0.98$.

Figure 3.2: *The relationship between the parameter r, of the Normal copula, and $\bar{r}_C$, for discrete variables with equal (a), and "complementary" (b) marginal distributions.*

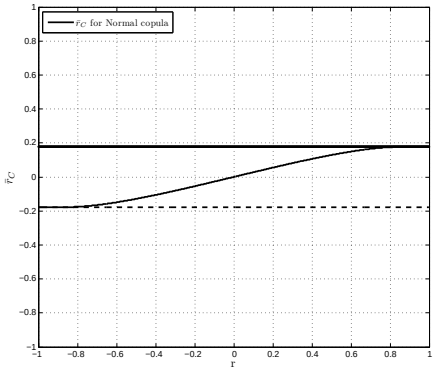


(a) $p_{1+} = p_{+1} = p_{2+} = p_{+2} = 0.01, p_{3+} = p_{+3} = 0.98$.

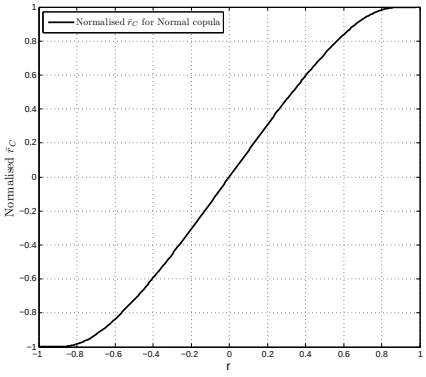


(b) $p_{2+} = p_{3+} = p_{+1} = p_{+2} = 0.01, p_{1+} = p_{+3} = 0.98$.

Figure 3.3: *The relationship between the parameter r, of Mardia's copula, and $\bar{r}_C$, for discrete variables with equal (a), and "complementary" (b) marginal distributions.*



(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 3.4: *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).*

we can calculate bounds for $r_{\bar{C}}$:

$$\frac{3\left(\sum_{i,j=1}^{m-1}(p_{i+}+p_{(i+1)+})(p_{+j}+p_{+(j+1)})\max\left(0,\sum_{k=1}^{i}p_{k+}+\sum_{l=1}^{j}p_{+l}-1\right)-\sum_{i,j=1}^{m-1}p_{i+}p_{+j}\right)}{\sqrt{\left(1-\sum_{i=1}^{m}p_{i+}^3\right)\cdot\left(1-\sum_{j=1}^{n}p_{+j}^3\right)}}$$
$$\leq r_{\bar{C}} \leq$$
$$\frac{3\left(\sum_{i,j=1}^{m-1}(p_{i+}+p_{(i+1)+})(p_{+j}+p_{+(j+1)})\min\left(\sum_{k=1}^{i}p_{k+},\sum_{l=1}^{j}p_{+l}\right)-\sum_{i,j=1}^{m-1}p_{i+}p_{+j}\right)}{\sqrt{\left(1-\sum_{i=1}^{m}p_{i+}^3\right)\cdot\left(1-\sum_{j=1}^{n}p_{+j}^3\right)}}. \tag{3.4.6}$$

These bounds are shown in Figure 3.4a. Since the bounds can be calculated, we can normalise the rank coefficient $\bar{r}_C$, such that it covers the entire interval $[-1,1]$, whatever the marginal distributions. The result of such a normalisation, in a particular case, is displayed in Figure 3.4b.

There are still open issues related to this topic. One of them is to determine how sensitive the relationship between the rank correlation of two discrete variables and the rank correlation of their underlying uniforms, is to the choice of the copula, and to the choice of marginal distributions (when the construction from equation (3.4.1) is used). It would also be worth describing this relation in the case of not ordered copulae.

The class of discrete distributions that we obtain in this way will obviously depend on the choice of the copula and its properties. It would be interesting to study if all bivariate discrete distributions can be obtained using this construction. If so, then we can also investigate what copula will best fit a given distribution, and how would one chose a certain copula for a specific distribution.

Chapter 4

Mixed Non-Parametric Continuous & Discrete Bayesian Belief Nets with Applications

The purpose of this chapter is to illustrate the use of BBNs in decision support systems. If we enrich the technique described in Chapter 2, Section 2.3 with the theory presented in Chapter 3, the result is an approach to mixed non-parametric continuous & discrete BBNs.

This approach is already successfully applied in two large ongoing projects, *CATS* and *Beneris*. Further we describe, in general terms, the two models used in these projects. More details of the approach are discussed on a simplified version of the latter, in Section 4.2.[1].

It is worth mentioning that both projects use UniNet, the software application where the approach to mixed non-parametric continuous & discrete BBNs has been implemented. The main program features are presented in the Appendix 7.1. Suffice here to say that UniNet was developed to support the *CATS* project. The software will be shortly available free from http://dutiosc.twi.tudelft.nl/∼risk/, together with supporting scientific documentation.

4.1 Ongoing Applications

CATS stands for *Causal Model for Air Transport Safety*. It is a large scale application on risks in the aviation industry, currently under development. The project is commissioned by the Netherlands Ministry of Transport and Water Management.

Beneris is a project undertaken by the European Union. The name of the project stands for *Benefit and Risk* and it focuses on the analysis of health benefits and risks associated with food consumption.

[1]Section 4.2 is based on the article Hanea and Kurowicka (2008), "Mixed non-parametric continuous and discrete bayesian belief nets", published in Advances in Mathematical Modeling for Reliability ISBN 978-1-58603-865-6 (IOS Press).

4.1.1 Causal Model for Air Transport Safety

The Netherlands Ministry of Transport and Water Management has commissioned a project for the realization of a causal model to be used for comparing alternatives for strengthening safety measures, for finding causes of incidents and accidents, and for quantification of the probability of adverse events in the aviation system (Ale et al. 2006; Morales-Napoles et al. 2007). The model so far covers the flight phases *take-off*, *en route*, and *approach-and-landing*. It involves probabilistic nodes whose marginal distributions are, in most cases retrieved from field data. In a few cases structured expert judgment is applied. The influences between probabilistic nodes are also elicited with a structured protocol described in (Morales et al. 2007). In addition to probabilistic nodes, the model also contains functional nodes that capture event sequence diagram[2], and fault tree modeling via Boolean functions. The current version of the model involves 644 discrete and continuous probabilistic nodes and 715 functional nodes. The model is pictured below. Neither the graphic resolution, nor the purpose of this chapter permits a detailed picture of the individual nodes.

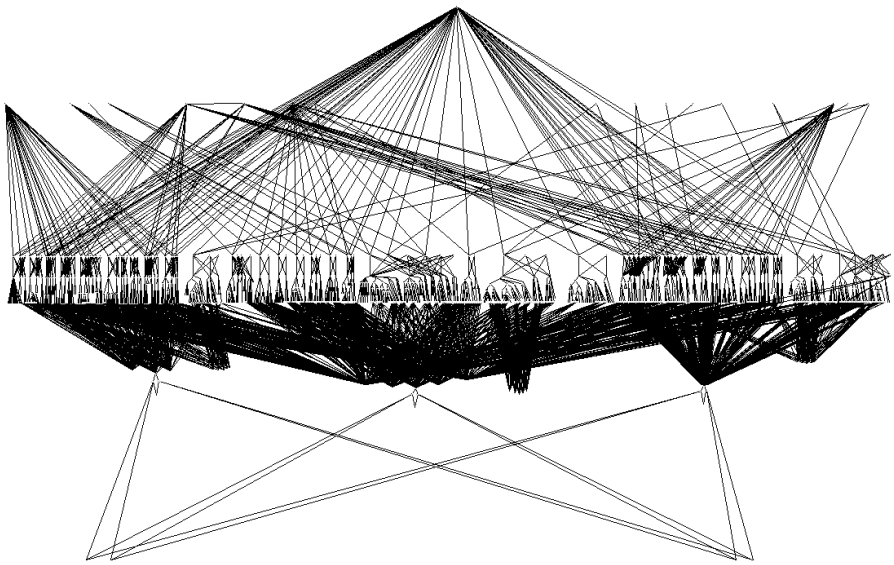


Figure 4.1: *BBN for CATS.*

The probabilistic nodes are the ones from the bottom half (until the middle "strip" of nodes) of the model from Figure 4.1. Events initiating accident scenarios in the CATS model are, to a large extent, a result of incorrect performance of humans.

[2]An event sequence diagram is a representation of an event tree which distinguishes different types of events.

The probabilistic nodes are measurable variables which influence human error probabilities. Sub-models for the probability of human errors have been developed for *flight crew* and *air traffic controller*. The flight crew performance model is partitioned over the three flight phases: take-off, en route, and approach-and-landing. It contains variables like: first officer/captain experience, first officer/captain training , fatigue, crew unsuitability, workload, etc. The air traffic controller performance model is also treated separately for the three flight phases. Some of the variables taken into account for the air traffic controller model are: traffic, man-machine interface, communication - coordination, air traffic controller experience, etc. The variables from these two sub-models account for the probabilistic relationships of the model. Another model for *maintenance crew performance* is under development, but not yet implemented at the moment of writing this thesis.

The nodes from the upper half of the BBN are functional nodes. The middle strip of nodes contains 31 clusters[3] of nodes that represent possible accident scenarios. Among these, one may find: aircraft system failure, air traffic controller event, aircraft handling by flight crew inappropriate, aircraft directional control related system failure, aircraft continues take off with contaminated wing, aircraft weight and balance outside limits during take-off, flight crew spatially disoriented, aircraft handling by flight crew during landing roll inappropriate, etc. The nodes one level up are more generic accident scenarios, acting like summary nodes. Some of these nodes represent: loss of control in flight, fire in flight, engine failure, runway overrun, collision with ground, in flight break up, air craft lands off runway. The topmost node represents the probability of having an incident or accident per flight[4]. This functional node is a combination of several specific accident scenarios. In principle, this node could be a combination of some generic scenarios from the upper level. However, this is not visible in the model.

It takes 2.67 minutes to sample the *CATS* model in UniNet[5]. The time to propagate evidence thorough the model is very close to the sampling time, e.g. it takes 2.25 minutes to resample the joint distribution conditioned on 3 variables.

4.1.2 Benefits and Risks

Beneris focuses on the analysis of health benefits and risks associated with food consumption[6]. Fish consumption is the first case study in the project (Jesionek and Cooke 2007). Its main goal is to estimate the health effects in a specified population as a result of exposure to various contaminants and nutrients through ingestion of fish. The population consists of the following age sub-groups: from 0

[3]Each cluster corresponds to an event sequence diagram, originally present in the model (Ale et al. 2006).

[4]Incidents or accidents are understood according to the International Civil Aviation Organization's definition, i.e. as an unintended event that causes death, injury, environmental or material damage.

[5]The default number of samples is 32.500. It takes 6.87 minutes to produce 100.000 samples.

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

to 2 years, from 2 to 18 years, from 18 to 55 years and older than 55 years. The latest version of the model is presented in Figure 4.2.

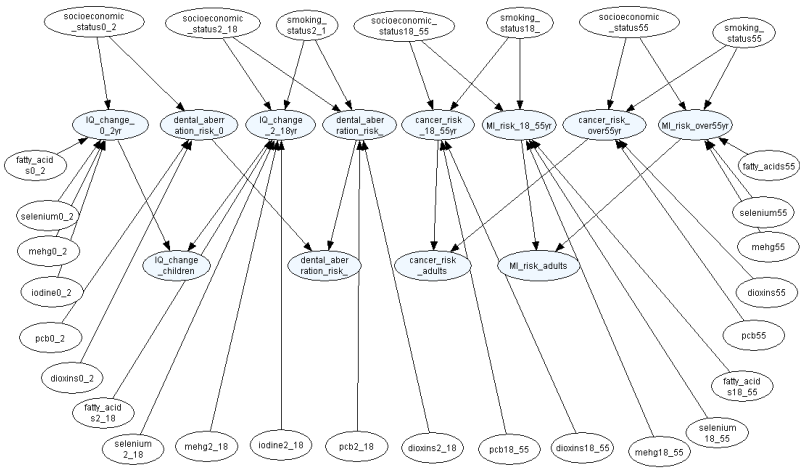


Figure 4.2: *BBN for Beneris.*

Some of the fish constituents of interest are: dioxins and furans, polychlorinated biphenyls (pcb), methyl mercury (MeHg), selenium, iodine, fish oils, etc. They are measured as yearly intake. Some of these factors (e.g. dioxins - furans, polychlorinated biphenyls) are persistent and bio-accumulative toxins which cause cancer in humans. Fish oil, on the other hand 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 and socioeconomic status are also taken into account. These factors are specific for each single age group. Smoking is measured as yearly intake of nicotine during smoking and passive smoking, while the socioeconomic status is measured by the number of years of schooling received, or to be received.

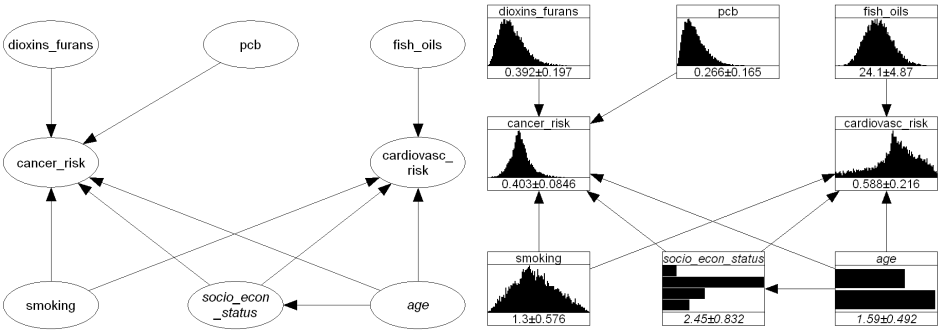
The health endpoints resulting from exposure to fish constituents are cancer, dental aberration, learning disability, and myocardial infarction. These health effects are defined in terms of remaining lifetime risks. Learning disability, however, is a more specific health effect. It is measured as a change in the IQ score relative to a baseline IQ[7]. All health endpoints considered in the model are influenced by (functions of) various parameters of fish constituents and personal factors.

[7]The model is under continuous development. The definitions of the variables can change in further stages of the project.

4.2 Highly Simplified Beneris

We will further discuss the approach to mixed non-parametric continuous & discrete BBNs using an example that is loosely based on the *Beneris* model presented in the previous section. We will only consider probabilistic nodes, since this is the theoretical background that we have explored in this thesis.

Figure 4.3a resembles the version of the model that we are considering for purely illustrative purposes. The variables of interest, for this version of the



(a) Simplified fish consumption BBN.

(b) Simplified fish consumption BBN with histograms.

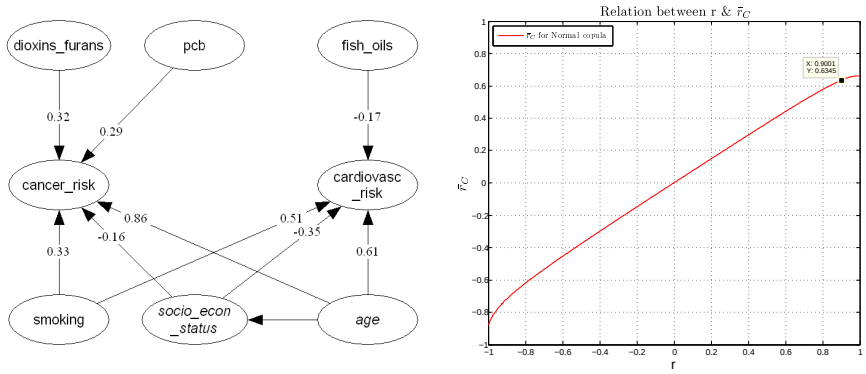
Figure 4.3: *Simplified Bayesian Belief Net for fish consumption risks.*

model, are cancer and cardiovascular risk, defined in terms of remaining lifetime risks.

We consider only 3 fish constituents, namely dioxins/furans, polychlorinated biphenyls, and fish oil. Smoking, and socioeconomic status are kept in the analysis. Instead of dividing the population in age sub-groups, we introduce an extra variable, age, 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 socioeconomic status is measured in this example by income, which is represented by a discrete variable with 4 income classes.

The distributions of the variables that form the BBN are presented in Figure 4.3b. They are chosen by the author 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. 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 author. Figure 4.4a 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 4.4: *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 chapter, 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 4.4b. 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 2.3, and calculate the sample rank correlation matrix (see Table4.1).

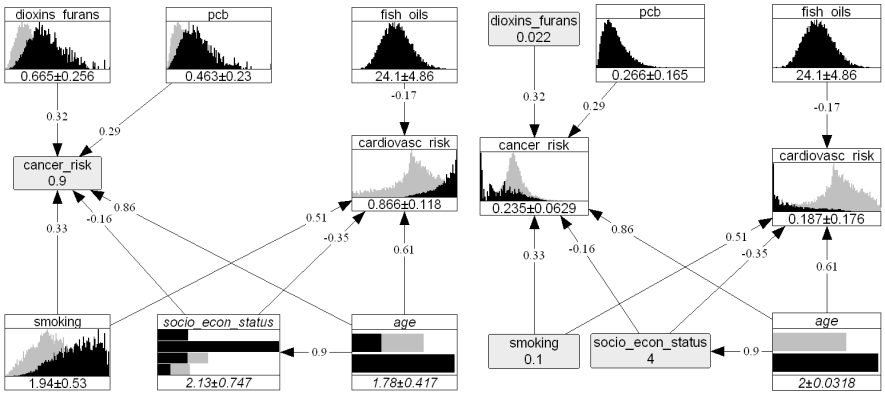
	dioxins furans	pcb	fish oils	smoking	socioecon. status	age	cancer risk	cardiovasc. risk
dioxins/furans	1	-0.0002	-0.0021	0.0012	0.0013	0.0012	0.322	0.0014
pcb	-0.0002	1	0.0033	0.0008	-0.0015	-0.0011	0.2718	-0.001
fish oils	-0.0021	0.0033	1	0.0015	-0.0007	-0.0022	-0.0006	-0.1654
smoking	0.0012	0.0008	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

Table 4.1: *The sample rank correlation matrix.*

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) using the relation from Remark 3.4.1.

We will further examine the situation in which there is a very high risk of cancer. We conditionalise on the 0.9 value of cancer risk and study in what way the other variables in the graph are affected by this information.



(a) Conditionalised BBN for cancer_risk = 0.9. **(b)** Conditionalised BBN for dioxins_furans=0.022, smoking=0.1, socio_econ_status=4.

Figure 4.5: *Diagnostic & predictive reasoning using the BBN.*

Figure 4.5a 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 much, and ingests a large amount of 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. In this case the BBN is used for diagnosis.

The conditionalisation in a BBN can be also used for prediction. For example one can be interested in the cancer risk of a person that inhales a very small amount of nicotine, has a high socioeconomic status and ingests very little dioxins and furans. Figure 4.5b presents how this information propagates through the graph. In this combination of factors, the expected value of the cancer risk decreases from 0.4 to 0.23. A substantial decrease can be also noticed in the cardiovascular risk. Because socioeconomic status and age are positively correlated, a high socioeconomic status results in a reduction of the population to the segment older than 35 years.

As in the case of the rank correlation between two continuous variables, a value for $\bar{r}_C$ can be either obtained from data, or from experts. The technique for eliciting (conditional) rank correlations for discrete variables is still an open issue.

Chapter 5

Mining and Visualising Ordinal Data with Non-Parametric Continuous BBNs[1]

5.1 Introduction

We shall further consider non-parametric BBNs from a completely different point of view, namely as a tool for mining ordinal multivariate data. Data mining is the process of extracting and analysing information from large databases. BBNs serve as a suitable framework for this purpose. The patterns of influence among variables can be represented as arcs in a BBN. Our aim is to learn the structure of a BBN that captures most of the dependencies present in a database. Specifying the structure of the model is one of the most important design choices in graphical modelling. Notwithstanding their potential, there are only a limited number of applications of graphical models on very complex and large databases.

An ordinal multivariate data set is one in which the numerical ordering of values for each variable is meaningful. A database of street addresses is not ordinal, but a database of fine particulate concentrations at various measuring stations is ordinal; higher concentrations are harmful to human health[2]. We describe a method for mining ordinal multivariate data using non-parametric BBNs, and illustrate this with ordinal data of pollutants emissions and fine particulate concentrations. The data are gathered from electricity generating stations and from collection sites in the United States over the course of seven years (1999 - 2005). The data base contains monthly emissions of SO_2[3] and NO_x[4] at different lo-

[1] This chapter is based on the paper Hanea et al. (2007), "Ordinal Data Mining with Non-Parametric Continuous Bayesian Belief Nets", accepted for publication in Computational Statistics and Data Analysis.

[2] Fine particulate concentration is measured on "at least an ordinal scale", i.e. the group of invariance transformations of its scale is a subgroup of the monotone increasing transformations.

[3] Sulfur dioxide is the chemical compound with the formula SO_2. This gas is the main product from the combustion of sulfur compounds and is of significant environmental concern.

[4] NO_x is a generic term for mono-nitrogen oxides (NO and NO_2). These oxides are produced during combustion, especially combustion at high temperatures.

cations, and monthly means of the readings of $PM_{2.5}$ concentrations at various monitoring sites. The notation $PM_{2.5}$ is used to describe particles of 2.5 micrometers or less in diameter. There are 786 emission stations[5] and 801 collector sites. This data set allows us to relate the emissions with the air quality and interpret this relationship.

Let us assume that we are interested in the air quality in Washington DC and how is this influenced by selected power plant emissions (see Figure 5.1). Additional variables that influence the $PM_{2.5}$ concentration in Washington DC are the meteorological conditions. We incorporate in our analysis the monthly average temperature, the average wind speed and wind direction.

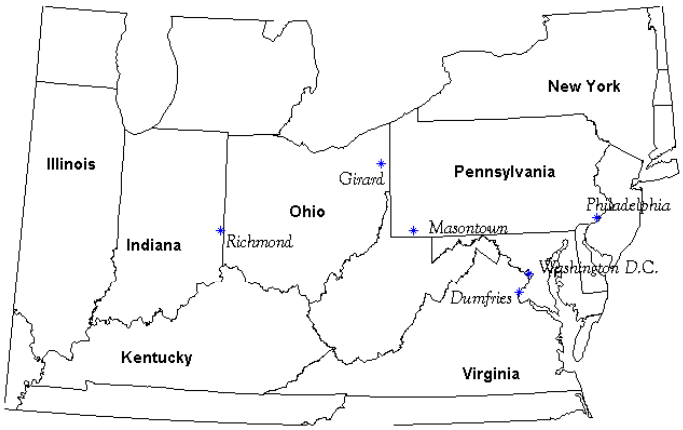


Figure 5.1: *Selected power plant emissions.*

A BBN for Washington DC ambient $PM_{2.5}$ is shown in Figure 5.2. This model is similar to the one described and analysed in Morgenstern et al. (2008). It involves the same 14 variables as nodes, but the arcs between them are different. There are 5 emission stations in the following locations: Richmond, Masontown, Dumfries, Girard and Philadelphia. For each such station, there are 2 nodes in the BBN. One corresponds to the emission of SO_2, and the other to the emission of NO_x. The variable of interest is the $PM_{2.5}$ concentration in Washington DC (DC_monthly_concPM25). There are 3 nodes that correspond to the meteorological conditions, namely the wind speed, wind direction and the temperature in DC. Conditional independence relations are given by the separation properties of the graph (see Section 5.4); thus nox_Philadelphia and DC_WindDir are independent conditional on DC_Temp and DC_WindSpeed. The methodology is designed specifically to handle large numbers of variables, in the order of several hundreds

[5]For most stations there is information on emissions of both SO_2 and NO_x, but for some we only have information about one or the other.

(see Morales-Napoles et al. (2007)), but a smaller number of variables is more suitable for explaining the method.

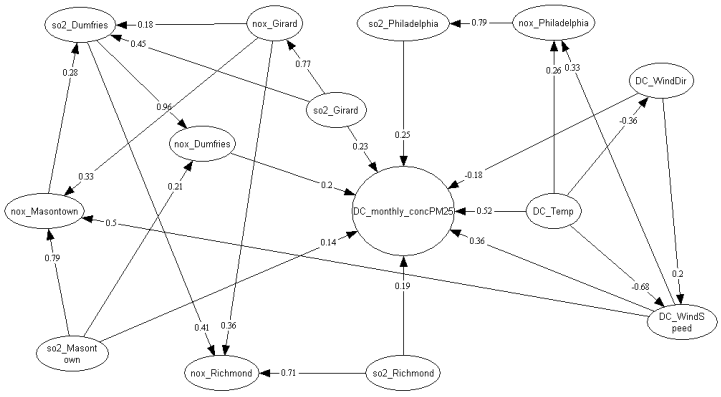


Figure 5.2: *BBN for Washington DC ambient $PM_{2.5}$.*

The most common methods to deal with continuous nodes are either to discretize them, or to assume joint normality. Both have disadvantages, as discussed in Chapter 1. In the former method, if all nodes are discretized to 10 possible values, a variable like DC_monthly_concPM25 (see Figure 5.2) would require a conditional probability table with 10^9 entries. Hence such models quickly become intractable. In the latter, the restriction to joint normality is rather severe. Figure 5.3 shows the same BBN as Figure 5.2, but the nodes are replaced by histograms showing the marginal distributions at each node. They are far from normal.

Our approach discharges the assumption of joint normality and builds a joint density for ordinal data using the joint normal copula. This means that we model the data as if it were transformed from a joint normal distribution. Influences are represented as conditional rank correlations according to the protocol explained in Chapter 2, Section 2.1. Other copulas could be used, but (to our knowledge) only the joint normal copula affords the advantages of rapid conditionalisation, while preserving the (conditional) independence for zero (conditional) correlation[6].

Rapid conditionalisation is perhaps the most important feature of a BBN from a user's standpoint. To illustrate, Figures 5.4 and 5.5 show the result of conditionalising the joint distribution on cold weather (275K) in Washington and low (Figure 5.4) and high (Figure 5.5) concentrations of $PM_{2.5}$ in Washington. The differences between the emitters' conditional distributions (black), and the original ones (gray), caused by changing the concentration, are striking, in spite of

[6]T-copula also allows for rapid conditionalisation but does not have the zero independence property.

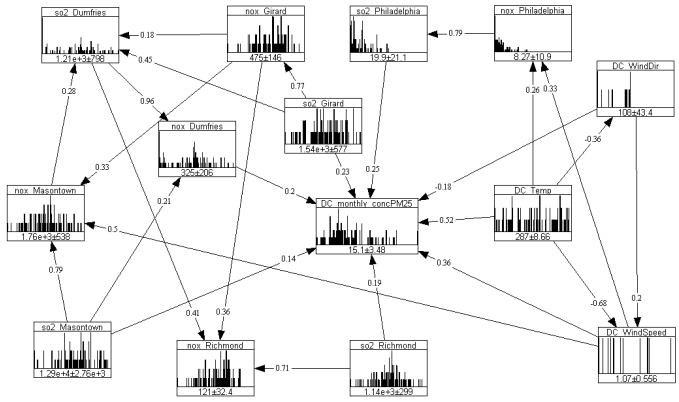


Figure 5.3: *Washington DC ambient $PM_{2.5}$ BBN with histograms.*

the relatively weak correlations with Washington's concentrations. Of course, rapid computations are of little value if the model itself cannot be validated. Validation involves two steps:

1. Validating that the joint normal copula adequately represents the multivariate data, and

2. Validating that the BBN with its conditional independence relations is an adequate model of the saturated graph.

Validation requires an overall measure of multivariate dependence on which statistical tests can be based. The discussion in Section 5.2.2 leads to the choice of the determinant of the correlation matrix as an overall dependence measure. This determinant attains the maximal value of 1 if all variables are uncorrelated, and attains a minimum value of 0 if there is linear dependence between the variables. We briefly sketch the two validation steps for the present example. Since we are dealing with copulae models, it is more natural to work with the determinant of the rank correlation matrices.

If we convert the original data to ranks and compute the determinant of the empirical rank correlation matrix (DER) we find the value 0.1518E-04. To represent the data with a joint normal copula, we must transform the marginals to standard normals, compute the correlation matrix, and compute the determinant of the normal rank correlation matrix (DNR) using the Pearson's transformation (see Chapter 1). This relation of correlation and rank correlation is specific to the normal distribution and reflects the normal copula. DNR is not in general equal to DER. In this case DNR = 0.4506E-04. Use of the normal copula typically introduces some smoothing into the empirical joint distribution, and this is

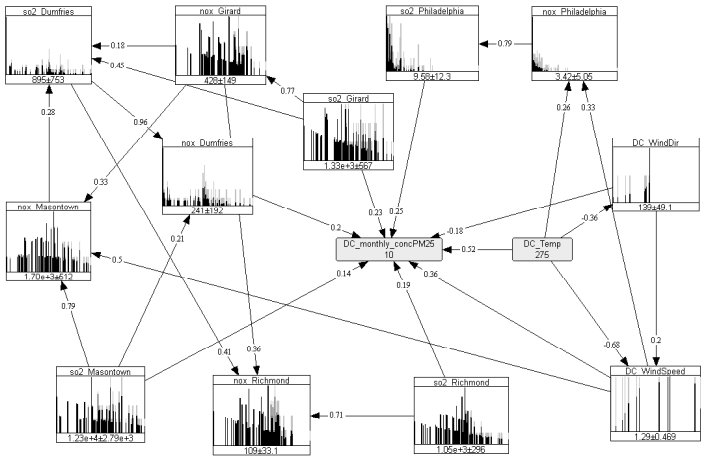


Figure 5.4: *Conditionalisation on low concentration of $PM_{2.5}$ for Washington DC and cold weather.*

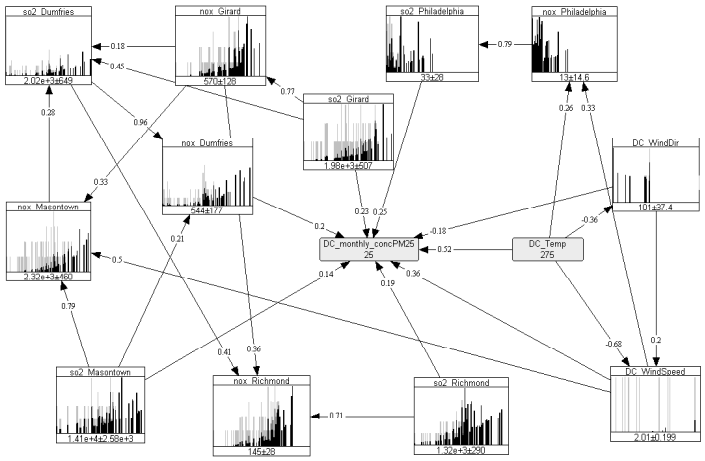


Figure 5.5: *Conditionalisation on high concentration of $PM_{2.5}$ for Washington DC and cold weather.*

reflected in a somewhat higher value of the determinant of the rank correlation matrix.

We can test the hypothesis whether this empirical rank distribution came from a joint normal copula in a straight forward way. We determine the sampling distribution of the DNR by simulation. Based on 1000 simulations, we find that the 90% central confidence interval for DNR is [0.0601E-04, 0.4792E-04]. The hypothesis that the data were generated from the joint normal copula would not be rejected at the 5% level.

DNR corresponds to the determinant of the saturated BBN, in which each variable is connected with every other variable. With 14 variables, there are 91 arcs in the saturated graph. Many of these influences are very small and reflect sample jitter. To build a perspicuous model we should eliminate noisy influences.

The BBN of Figure 5.2 has 26 arcs. To determine whether these 26 arcs are sufficient to represent the saturated graph, we compute the determinant of the rank correlation matrix based on the BBN (DBBN). This differs from DNR, as we have changed many correlations to zero and introduced conditional independencies. In this case, DBBN = 1.5092E-04. We determine the sampling distribution of the DBBN by simulation. Based on 1000 simulations, we find that the 90% central confidence interval for DBBN is [0.2070E-04, 1.5905E-04]. DNR is within the above mentioned 90% central confidence band. A simpler BBN involving only 22 arcs is shown in Figure 5.6. It has a DBBN of 4.8522E-04. The 90% central confidence interval for this DBBN is [0.7021E-04, 5.0123E-04]. This interval does not contain DNR and would be rejected.

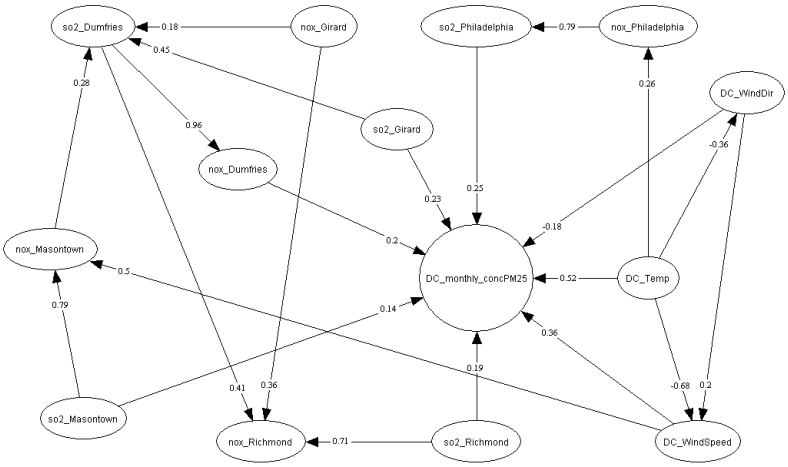


Figure 5.6: *Simplified BBN with 22 arcs.*

In general, changing correlations disturbs the positive definiteness of the rank correlation matrix. Moreover, the nodes connected in a BBN represent only a portion of the correlations. We can apply simple heuristics to search for a suitable BBN model without becoming embroiled in matrix completion and positive definitness preservation because of the way we represent joint distributions in a BBN. The conditional rank correlations in a BBN are algebraically independent and, together with the graphical structure and marginal distributions, uniquely determine the joint distribution. These facts have been established in Chapter 2, Section 2.1. The key idea is to link a BBN with a nested sequence of regular vines.

The joint distribution of a set of variables can be graphically represented as a BBN or as a vine, in an equivalent way. Both in the BBN and in the vine, one will have to specify the (conditional) rank correlations associated with the arcs/edges. In some cases these two structures require exactly the same (conditional) rank correlations. But, as we saw in Chapter 2, this is not always the case. If a (conditional) rank correlation specification is available for the arcs of a BBN, this can be translated to a specification for the vine[7]. In this process additional computations may be required. For arbitrary choice of copula this can constitute a big disadvantage in terms of computational complexity. However for the normal copula this disadvantage vanishes, as we can always recalculate required correlations for a given ordering of the variables.

Some important properties of vines translate almost immediately to corresponding properties of non-parametric BBNs, for example the result of Theorem 1.2.3. We can formulate and prove an analog property for BBNs, which plays the central role in model inference from Section 5.2.

Theorem 5.1.1. *Let D be the determinant of an n-dimensional correlation matrix ($D > 0$). For any partial correlation BBN specification*

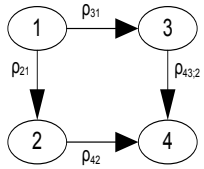
$$D = \prod \left(1 - \rho^2_{ij;D_{ij}}\right),$$

where $\rho_{ij;D_{ij}}$ is the partial correlation associated with the arc between node i and node j, D_{ij} is the conditioning set for the arc between node i and node j, and the product is taken over all arcs in the BBN.

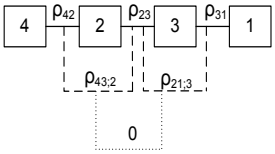
Proof. To prove this fact we will use the connection between BBNs and vines. If the BBN can be represented as a vine with the same partial correlation specification on its edges, the result follows from Theorem 1.2.3. If this is not the case, namely if the partial correlation specification for the vine differs from the one for the BBN, we will use the equation from Theorem 1.2.3 sequentially. Let us assume that we have a sampling order for the variables. Without loss of generality we may consider this order as being $1, 2, ...n$. We will construct a vine for these

[7]This is true when using non-constant conditional copulae (hence non-constant conditional correlations). In the case of normal copula, it is also true for constant conditional correlations.

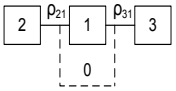
variables which contains: variable n, the parents of variable n, and the variables independent of n given its parents in the BBN (in this order). For example, let us consider the BBN from Figure 5.7a. A sampling order of these variables is 1,2,3,4. The D-vine corresponding to this BBN is shown in Figure 5.7b.



(a) Partial correlation specification BBN on 4 variables.



(b) Partial correlation specification vine on 4 variables.



(c) Partial correlation specification vine on the last 3 variables.

Figure 5.7: *The connection between a partial correlation BBN specification and a partial correlation vine specification.*

The BBN and the vine constructed as above will have the same correlation matrix. The determinant of the correlation matrix can be calculated using Theorem 1.2.3. The construction of the vine in this specific way, ensures that the non-zero partial correlations that have variable X_4 in the conditioned set, i.e. ρ_{42}, $\rho_{43;2}$, are the same as the ones associated to the arcs between X_2 and X_4, and between X_3 and X_4. Using the conditional independence statements on the BBN and the normal copula, we know $\rho_{41;23} = 0$.

In the general case, the non-zero partial correlations that have variable n in the conditioned set correspond to partial correlations associated to the arcs of the BBN that connect n with its parents. Therefore, the determinant of the correlation matrix will be a product that contains 1 minus these partial correlations squared. The rest of the terms in this product correspond to the determinant of the correlation matrix of first $n-1$ variables. For the particular case above:

$$D = \left(1-\rho_{42}^2\right)\left(1-\rho_{43;2}^2\right)\left[\left(1-\rho_{23}^2\right)\left(1-\rho_{31}^2\right)\left(1-\rho_{21;3}^2\right)\right].$$

The product of the last 3 terms corresponds to the determinant of the correlation matrix of X_1, X_2, X_3. We can now reorder the variables and construct the vine from Figure 5.7c. Then:

$$\left(1-\rho_{23}^2\right)\left(1-\rho_{31}^2\right)\left(1-\rho_{21;3}^2\right) = \left(1-\rho_{21}^2\right)\left(1-\rho_{31}^2\right)\left(1-\rho_{32;1}^2\right) = \left(1-\rho_{21}^2\right)\left(1-\rho_{31}^2\right).$$

For any regular vine on $n-1$ variables, the product of 1 minus squared partial correlations assigned to the edges of the vine is the same, hence we can reorder the variables such that they will correspond to the ones from the edges of the BBN. If this is not possible for the entire vine on $n-1$ variables, we repeat the previous step sequentially. □

The partial correlation BBN represents a factorisation of the determinant of the correlation matrix. This property is crucial in our algorithm for learning the structure of a BBN.

The rest of this chapter is organised as follows: in Section 5.2.1 we present a short overview of the existing methods for learning BBN structures from data. In order to introduce our approach we need to select a measure of multivariate dependence. Section 5.2.2 contains a discussion about various such measures. In Section 5.2.3 we introduce our learning algorithm, and in Section 5.3 we present this approach using the database of pollutants emissions and fine particulate concentrations. In the last part of this chapter we discuss alternative ways to calculate the correlation matrix of a BBN and illustrate how these may speed up the updating algorithm.

5.2 Learning the Structure of a BBN

5.2.1 Overview of Existing Methods

Data mining is the process of extracting and analysing information from large databases. For discrete data, BBNs are often used as they describe joint distributions in an intuitive way and allow rapid conditionalisation (Cowell et al. 1999).

In the process of learning a BBN from data, two aspects can be of interest: learning the parameters of the BBN, given the structure, and learning the structure itself.

Most of the current methods to learn the structure of a BBN focus on discrete or Gaussian variables (Spirtes et al. 1993). There are two main classes of algorithms for learning the structure of a BBN. One class *scores* a BBN structure based on how well it fits the data, and attempts to produce one that optimises the score. A score function is used to choose the best model within the group of all possible models for the network. This poses very difficult problems since the space of all possible structures is at least exponential in the number of variables. Therefore computing the score of every BBN structure is not possible in all but the most trivial domains. Instead, heuristic search algorithms are used in practice (Lam and Bacchus 1994; Heckerman 1995).

The alternative approach uses constraints such as independence relations present in the data, to reconstruct the structure. A number of statistical conditional independence tests are conducted on the data, and their results are used to make inferences about the structure (Cheng et al. 1997; Spirtes et al. 1993; Pearl and Verma 1991).

Although many of these algorithms provide good results on some small data sets, there are still several problems. One of these problems is that many algorithms require additional information, for example an ordering of the nodes to reduce the search space (see Cooper and Herskovits 1992; Heckerman et al. 1997; Cheng et al. 1997). Unfortunately, this information is not always available.

To our knowledge, the few methods that can handle non-parametric continuous variables (e.g., Margaritis 2005) can hardly be applied in domains with a large number of variables that are densely connected. Moreover the existing structure learning algorithms are slow, both in theory (Chickering et al. 1994) and in practice e.g., most constraint-based algorithms require an exponential number of conditional independence tests.

This motivates us to develop an algorithm for learning a BBN structure from data, which is more suitable for real world applications. Our goal is to learn the structure from an ordinal multivariate data set that may contain a large number of variables. This learning algorithm will not make any assumptions about the marginal distributions of the variables. We want to be able to learn such structures fast and use it further, for prediction purposes.

Comparisons of our method for learning the structure of a non-parametric continuous BBN with other existing methods are difficult to conduct for several reasons. To our knowledge, in most of the learning algorithms there are two approaches to deal with continuous variables. One is to assume that the variables belong to a family of parametric distributions (e.g. Heckerman and Geiger 1995; John and Langley 1995), and the other one is to use the discretised version of the variables (e.g. Friedman and Goldszmidt 1996). We use neither of the two methods.

An algorithm that deals with non-parametric continuous variables is proposed in Margaritis (2005). The authors of Margaritis (2005) develop a conditional independence test for continuous variables, which can be used by any existing independence-based BBN structure learning algorithm. The method is evaluated on two real-world data sets: BOSTON-HOUSING and ABALONE using the PC algorithm (Spirtes et al. 1993). We investigate the structure obtained for the BOSTON-HOUSING data set. This data set is available at http://archive.ics.uci.edu/ml/datasets.html. The data concerns housing values in suburbs of Boston. It contains 14 variables (13 continuous variables and a binary one) and 506 samples. In the structure presented in Margaritis (2005), the variable ZN is independent of all the others. If we calculate the empirical rank correlation matrix we find that ZN is correlated with other variables with high correlations, e.g.: 0.615, -0.643, -0.635. This result indicates that the method is inadequate and obviates further comparison.

5.2.2 Multivariate Dependence Measures

Inferring the structure of a BBN from data requires a suitable measure of multivariate dependence. Multivariate dependence measures are discussed in Micheas and Zografos (2006, Joe (1990, Schmid and Schmidt (2007). In Micheas and Zografos (2006) Renyi's axioms (Renyi 1959) for bivariate dependence are extended for the multivariate case and some representation results are proven.

Although multivariate dependence measures are not the focus of the present study, it is convenient to motivate the choice of such a measure by reference to a

set of axioms similar to that of Micheas and Zografos (2006).

We propose a set of axioms that specify properties of a multivariate dependence measure. It is convenient to restrict such measures to the $[0, 1]$ interval, with 1 corresponding to independence. $D_{1,...,n}$ denotes such measure. $\ell(X_1, ..., X_n)$ denotes the linear span of the variables $(X_1, ..., X_n)$. That is the set of vectors which can be written as affine combinations of $(X_1, ..., X_n)$. $n!$ is the set of all permutations of $\{1, ..., n\}$; π is a permutation from $n!$; $\perp \{1, ..., n\}$ says that the variables $(X_1, ..., X_n)$ are independent; $f_{1,...,n}$ is the density of $(X_1, ..., X_n)$ and f_i is the density of X_i.

We propose the following axioms:

AX 1 $0 \leq D_{1,...,n} \leq 1$;

AX 2 $\forall i$, $D_i := 1$;

AX 3 $\forall \pi \in n!$, $D_{1,...,n} = D_{\pi(1),...,\pi(n)}$;

AX 4 $K, J \subseteq \{1, ..., n\}$, $K \cap L = \emptyset$, $X_K \perp X_J \Longrightarrow D_{K,J} = D_K D_J$;

AX 5.1 $\perp \{1, ..n\} \Longrightarrow D_{1,...,n} = 1$;

AX 5.2 $\perp \{1, ..n\} \Longleftrightarrow D_{1,...,n} = 1$;

AX 6.1 $X_1 \in \ell(X_2, ..., X_n) \Longrightarrow D_{1,...,n} = 0$;

AX 6.2 $D_{1,...,n} = 0$, $D_{2,...,n} > 0 \Longrightarrow X_1 \in \ell(X_2, ..., X_n)$;

AX 7.1 $X_1 = g(X_2, ..., X_n)$ on a set of positive measure, where g is a measurable function $\Longrightarrow D_{1,...,n} = 0$;

AX 7.2 $D_{1,...,n} = 0$, $D_{2,...,n} > 0 \Longrightarrow X_1 = g(X_2, ..., X_n)$ on some set of positive measure.

We define a conditional dependence measure as:

$$D_{1,...,k;k+1,...,n} = \frac{D_{1,...,n}}{D_{k+1,...,n}}, \qquad D_{k+1,...,n} > 0.$$

Evidently

$$D_{1,...,n} = D_{1;2,...,n} D_{2;3,...,n} ... D_{n-1;n};$$

where $D_{n-1;n} = D_{n-1n}$ and we can specify a dependence measure by specifying the conditional dependence measures.

We note that AX 4 is stronger than its corresponding axiom in Micheas and Zografos (2006), and AX 7.1 & AX 7.2 are a bit weaker than their counterpart in Micheas and Zografos (2006). Axioms 6.1 and 6.2 explicitly capture the notion of linear dependence.

Proposition 5.2.1. *$D_{1,...,n} = Det(C)$, with C the correlation matrix of $X_1, ..., X_n$ satisfies AX 1, AX 2, AX 3, AX 4, AX 5.1, AX 6.1, AX 6.2.*

Proof. Let $D_{1,...,n} = Det(C)$. The first three axioms and AX 5.1 are obvious. For AX 4, suppose the correlation matrix has diagonal blocks $C_{1,...,k}$ and $C_{k+1,...,n}$. Let $C_{1,...,k} \oplus \mathbf{1}(k+1,...,n)$ denote the $n \times n$ matrix whose first $k \times k$ cells are $C_{1,...,k}$, whose diagonal entries $k+1,...,n$ are 1's, and whose other cells are 0's. Similarly, let $\mathbf{1}(1,...,k) \oplus C_{k+1,...,n}$ denote the matrix whose first k diagonal entries are 1's, whose last $k+1,...,n$ entries are $C_{k+1,...,n}$, and whose other cells are 0's. Then:

$$\begin{aligned} Det(C) &= Det\left(C_{1,...,k} \oplus \mathbf{1}(k+1,...,n) \times \mathbf{1}(1,...,k) \oplus C_{k+1,...,n}\right) \\ &= Det\left(C_{1,...,k} \oplus \mathbf{1}(k+1,...,n)\right) \times Det\left(\mathbf{1}(1,...,k) \oplus C_{k+1,...,n}\right) \\ &= Det(C_{1,...,k}) \times Det(C_{k+1,...,n}) = D_{1,...,k} D_{k+1,...,n}. \end{aligned}$$

To prove that axioms 6.1 and 6.2 hold, we use equation 1.2.2. $D_{1,...,n}$ is zero if and only if at least one of the multiple correlations in equation 1.2.2 is 1. If $D_{2,...n} > 0$, then $R^2_{1:2,...,n} = 1$, which means that X_1 is an affine combination of $(X_2,...,X_n)$. □

We will further discuss two other multivariate dependence measures. In order to introduce the first one we will first define the concept of *mutual information.*

Definition 5.2.1. *Let f and g be densities on $\mathbb{R}^n$, with f absolutely continuous with respect to g;*

- *the **relative information** of f with respect to g is:*

$$I(f|g) = \int_1 ... \int_n f(x_1,...,x_n) ln\left(\frac{f(x_1,...,x_n)}{g(x_1,...,x_n)}\right) dx_1...dx_n.$$

- *the **mutual information** of f is:*

$$MI(f) = I(f|\prod_{i=1}^{n} f_i),$$

If f is a joint normal density, then: $MI(f) = -\frac{1}{2}\log(D)$, where D is the determinant of the correlation matrix. This relation suggests that we can use $e^{-2MI(f)}$ as another multivariate dependence measure. $e^{-2MI(f)}$ satisfies AX 5.2. In Joe (1989) it is shown that $e^{-2MI(f)}$ satisfies AX 7.1 and 7.2 (AX 6.2 is not satisfied), moreover it is invariant under measurable and bijective transformations of each of the $X_i's$. Unfortunately, efficient methods for computing the sample MI are not available.

The multivariate Spearman's correlation (Schmid and Schmidt 2007) is sometimes proposed as a measure of multivariate dependence. For bivariate dependence, the Spearman's rank correlation is given by (Nelsen 1999):

$$r(X_1, X_2) = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3 = 12 \int_0^1 \int_0^1 uvc(u, v) du dv - 3. \quad (5.2.1)$$

where $C(u, v)$ is the copula for X_1,X_2, and $c(u, v)$ is the copula density. In higher dimensions the appropriate generalisations of the two integrals in equation 5.2.1 are not equal and a variety of possible generalisations exist (Schmid and Schmidt 2007). In three dimensions the version based on the copula density[8] reads (Schmid and Schmidt 2007):

$$r(X_1, X_2, X_3) = 8 \int_0^1 \int_0^1 \int_0^1 uvwc(u, v, w) du dv dw - 1.$$

Using the bivariate elliptical copula (Kurowicka, Misiewicz, and Cooke 2000) with a Markov multivariate copula that satisfies $c(u, v, w) = c(u, v)c(v, w)$, and using the fact that for the elliptical copula $E(U|V = v) = v\rho(U, V)$, (Kurowicka and Cooke 2006b), it follows that:

$$r(X_1, X_2, X_3) = 2\rho(U, V)\rho(V, W) - 1.$$

This entails that if $\rho(U, V) = 0$, then $r(X_1, X_2, X_3) = -1$, which is difficult to interpret.

On the basis of the above discussion we conclude that the determinant of the correlation matrix is a reasonable measure of multivariate dependence. In working with non-parametric BBNs, it is more convenient to focus on the multivariate dependence in the copula.

5.2.3 *Learning the Structure of a Non-Parametric Continuous BBN with the Normal Copula*

Suppose we have a multivariate data set. We may distinguish:

- DER = the determinant of the empirical rank correlation matrix;
- DNR = the determinant of the rank correlation matrix obtained by transforming the univariate distributions to standard normals, and then transforming the product moment correlations to rank correlations using Pearson's transformation (see Proposition 1.2.1);
- DBBN = the determinant of the rank correlation matrix of a BBN using the normal copula.

[8] The version based on the copula density is dedoted by ρ_2 in Schmid and Schmidt (2007).

DNR will generally differ from DER because DNR assumes the normal copula, which may differ from the empirical copula. A rough statistical test for the suitability of DNR for representing DER is to obtain the sampling distribution of DNR and check whether DER is within the 90% central confidence band of DNR. If DNR is not rejected on the basis of this test, we shall attempt to build a BBN which represents the DNR parsimoniously. Note that the saturated BBN will induce a joint distribution whose rank determinant is equal to DNR, since the BBN uses the normal copula. However, many of the influences only reflect sample jitter and we will eliminate them from the model.

Searching for a perspicuous model by eliminating arcs from the saturated graph is a data compression technique, and may be compared with other compression techniques. Factor analysis (Lawley and Maxwell 1963) for example seeks to express all variables as linear combinations of a smaller number of variables. Compression is accomplished by lowering the rank of the correlation matrix. The method of model selection presented in (Whittaker 1990), in contrast, seeks to eliminate influences between variables i and j when the partial correlation between them, given all other variables, is suitably small. In other words, the method from (Whittaker 1990) compresses by setting partial correlations of maximal order equal to zero. However the zeroing operation may perturb the positive definiteness of the correlation matrix. Both factor analysis and the method in (Whittaker 1990) assume a joint normal distribution. Here, the joint normality assumption is relaxed to the assumption of a normal copula. Setting partial correlations in a BBN equal to zero does not encounter the problem of positive definiteness, due to the connection between BBN's and regular vines.

If the BBN is not saturated, then DBBN $>$ DNR. We will use the result from Theorem 5.1.1 in building a BBN from data, in the context of the normal copula vine approach. Having a conditional rank correlation specification for the arcs of a BBN and using the normal copula, entails a partial correlation BBN specification. Moreover, the zero partial correlations will correspond to the conditional independence statements encoded in the BBN structure. We will build the BBN by adding arcs between variables only if the rank correlation between those two variables is among the largest. We will also remove arcs from the BBN, which correspond to very small rank correlations. The heuristic we are using is that partial correlations are approximately equal to conditional rank correlations. This is a reasonable approximation if we consider the following: we use the normal copula to realise the (conditional) rank correlations associated to the arcs of the BBN; the relation between (conditional) rank correlation and conditional (product moment) correlation is calculated using Pearson's transformation; for joint normal variables, the conditional (product moment) correlations and the partial correlations are equal.

The procedure for building a BBN to represent a given data set is not fully automated, as the directionality of (some of) the arcs will reflect causal or temporal relations which can never be extracted from data. The result of introducing

arcs to capture causal or temporal relations is called a Skeletal BBN. The general procedure can then be represented as:

1. Verify that DER is not outside the plausible central confidence band for DNR;
2. Construct a Skeletal BBN;
3. If DNR is within the 90% central confidence band of the determinant of the Skeletal BBN, then stop, else continue with the following steps;
4. Find the pair of variables (X_i, X_j) such that the arc (i, j) is not in the BBN and r_{ij}^2 is greater than the squared rank of any other pair not in the BBN. Add an arc between nodes i and j, and recompute DBBN together with its 90% central confidence band.
5. If DNR is within the 90% central confidence band of DBBN, then stop, else repeat step 4.

The 90% central confidence band may be replaced by the the 95% or 99% central confidence bands.

The resultant BBN may contain nodes that have more than one parent. If the correlations between the parents of a node are neglected in the BBN (i.e. if the parents are considered independent), then DBBN will be different for different orderings of the parents. These differences will be small if the neglected correlations are also small.

In general, there is no "best" model; the choice of directionality may be made on the basis of non-statistical reasoning. Some small influences may be included because the user wants to see these influences, even though they are small. There may be several distinct BBNs which approximate the saturated BBN equally well.

5.3 Ordinal $PM_{2.5}$ Data Mining with UniNet

We illustrate our method for learning a BBN from data using the ordinal multivariate data set that we briefly introduced in Section 5.1. The data are gathered from electricity generating stations and from collection sites in the United States over the course of seven years (1999 - 2005). The data base contains monthly emissions of SO_2 and NO_x in different locations and monthly means of the readings of $PM_{2.5}$ concentrations at various monitoring sites. Since we have monthly data over the course of seven years, the data set will contain 84 multivariate samples.

There are 786 emission stations and 801 collector sites. Meteorological information on temperature wind speed and wind direction is also available at, or near, all cites. Although the method is designed to handle large numbers of variables, we adopt a smaller set for purposes of illustration. We consider one collector at Washington D.C. temperature, wind speed and wind direction at Washington DC,

and emissions from five stations which are upwind, under prevailing winds, and emit large quantities of SO_2 and NO_x: Richmond, Masontown, Dumfries, Girard and Philadelphia (see Figure 5.8). The goal is to build a BBN that captures the dependence structure between these variables, using the approach presented in the previous section. This learning approach has been implemented into UniNet, hence all analysis and graphs are produced using UniNet.

Figure 5.8: *BBN on 14 nodes with no arcs.*

The distinctive feature of this approach is that we take the one dimensional marginal distributions directly from the data, and model the dependence with the joint normal copula. The hypothesis that the dependence structure in the data is that of a joint normal copula can be tested by the method described in Section 5.1. Once we have a suitable copula, we can condition any set of variables on values of any other set of variables.

Standard regression analysis also computes conditional distributions. For data sets like that encountered here, however, the BBN approach with the normal copula offers several advantages:

- We obtain the full conditional distribution, not just the mean and variance.
- We do not assume that the predicted variable has constant conditional variance, indeed the conditional distributions do not have constant variance.
- The emitters tend to be strongly correlated to each other and weakly correlated to the collectors, hence if we marginalize over a small set of upwind emitters, we have many "missing covariates" with strong correlations

to the included covariates. This will bias the estimates of the regression coefficients. The BBN method, in contrast, simply models a small set of variables, where other variables have been integrated out. There is no bias; the result of first marginalising then conditionalising is the same as first conditionalising then marginalising.

- The set of regressors may have individually weak correlations with the predicted variable, but may be collectively important. On small data sets, the confidence intervals for the regression coefficients may all contain zero and their collective importance would be missed.

The discussion in the previous section led to the choice of the determinant of the rank correlation matrix as an overall dependence measure. This determinant attains the maximal value of 1 if all variables are independent, and attains a minimum value of 0 if there is linear dependence between the ranked variables. Figures 5.9 and 5.10 compare the empirical rank correlation matrix with the normal rank correlation matrix. It can be noticed that the highest correlations are in the same positions in both matrices. Moreover all differences are of order of 10^{-2}. For 84 samples, the approximate upper critical values of Spearman's rank correlation, are given in the table below (Ramsey 1989).

Table 5.1: *Critical values of Spearman's rank correlation for 84 samples.*

N	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
84	0.181	0.215	0.254	0.280

The α values correspond to a one-tailed test of the null hypotheses that the rank correlation is 0.

Figure 5.8 shows the 14 variables, as nodes in a BBN with no arcs. Hence, we start by considering these variables as being independent. We obtain DBBN = 1. In general, if the BBN is not saturated, then DBBN > DNR. Following the general procedure presented in the previous section we start adding arcs between variables whose rank correlation (in the normal rank correlation matrix) are among the largest. By doing so, we decrease the value of DBBN. UNINET allows us to visualise the highest rank correlations (see Figure 5.10). We add 16 arcs to the BBN, most of which correspond to the highest rank correlations. Nevertheless, our interest is to quantify the relation between Washington DC and the rest of the variables involved, hence we also add arcs that carry information about their direct relationship. The resultant BBN is shown in Figure 5.11. UNINET calculates from data the (conditional) rank correlations that correspond to the arcs of the BBN.

The determinant of the rank correlation matrix based on the new BBN differs from DNR, as this BBN is not saturated, hence it contains conditional independencies that are not present in the data. In this case DBBN = 3.6838E-04 and its 90% central confidence interval is [0.5552E-04, 3.7000E-04]. We notice that

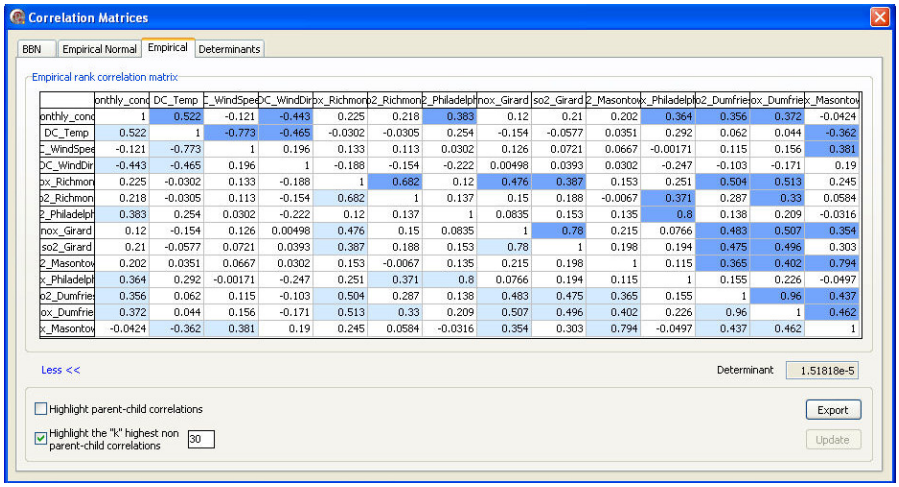


Figure 5.9: *The empirical rank correlation matrix of the 14-dimensional distribution.*

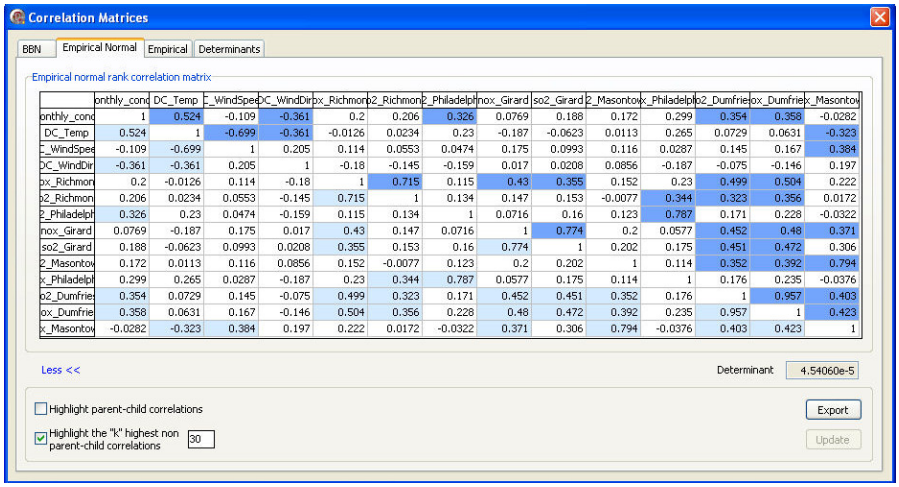


Figure 5.10: *The normal rank correlation matrix of the 14-dimensional distribution.*

DNR is not within this interval. In consequence, we need to add more arcs to the BBN. Following the same idea of quantifying direct influence on the air quality in Washington DC, we add 4 more arcs. The resultant BBN with 20 arcs is shown in Figure 5.12.

The 90% central confidence interval for the determinant of the rank correlation matrix based on the new BBN is [0.4617E-04, 2.6188E-04] and DNR is still outside this interval.

Adding arcs that do not necessarily correspond to the highest correlations might increase the number of iterations needed in order to obtain a valid structure. Moreover, the resultant BBN becomes more complicated. Nevertheless, there are situations in which we are more interested to represent certain direct influences between our variables, rather than obtaining a sparse structure.

We continue by adding arcs that correspond to the highest correlations from the matrix. We obtain the BBN from Figure 5.13.

The value of DBBN for the last BBN is 1.4048E-04. DNR falls inside the confidence interval for DBBN, which is [0.1720E-04, 1.6270E-04]. We conclude that this BBN with its conditional independence relations is an adequate model of the saturated graph.

We can continue looking for a more convenient representation (with less arcs) by changing very small correlations to zero, while disturbing the determinant as little as possible. We will now remove 4 arcs from the BBN (see Figure 5.14). DBBN changes to 1.5092E-04. This change is not significant, so the new BBN on 26 arcs is an adequate model as well. If we further reduce the number of arcs, we obtain the structure from Figure 5.6, whose determinant is 4.8522E-04, and would be rejected.

Another procedure for building a BBN to represent a given data set, would be to begin with the saturated graph[9], rather than with the empty one. The saturated BBN will induce a joint distribution whose rank determinant is equal to DNR, since the BBN uses the normal copula. Further we will remove those arcs that are associated with very small (close to zero) correlations, such that the value of DNR stays inside the confidence interval for DBBN.

It is worth mentioning that the BBN structure learned from the data set, using one approach or another, will not be unique. Adding/deleting different arcs from the BBN may provide a different suitable structure.

5.4 Alternative Ways to Calculate the Correlation Matrix of a BBN

In both, learning the structure of the BBN and the conditioning step, an important operation is calculating the correlation matrix from the partial correlations specified. To do so, we are repetitively using equation 1.2.1. When working with

[9]One of the possible saturated graphs. If the sampling order of the variables is known, this procedure might be more appropriate.

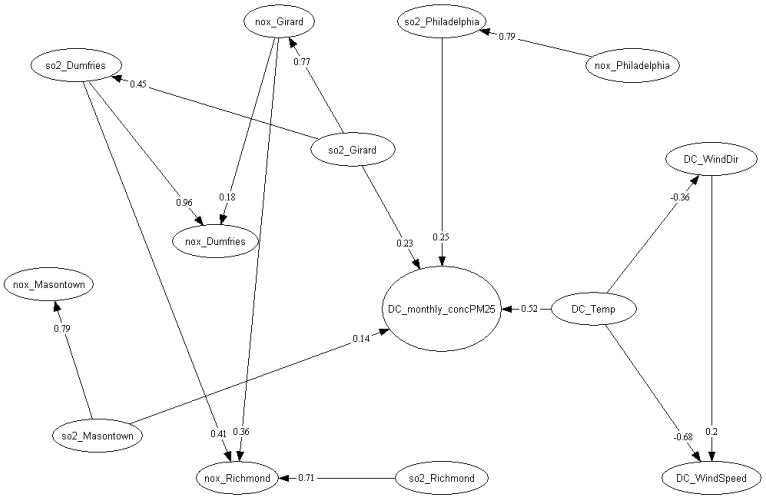


Figure 5.11: *BBN on 14 nodes with 16 arcs.*

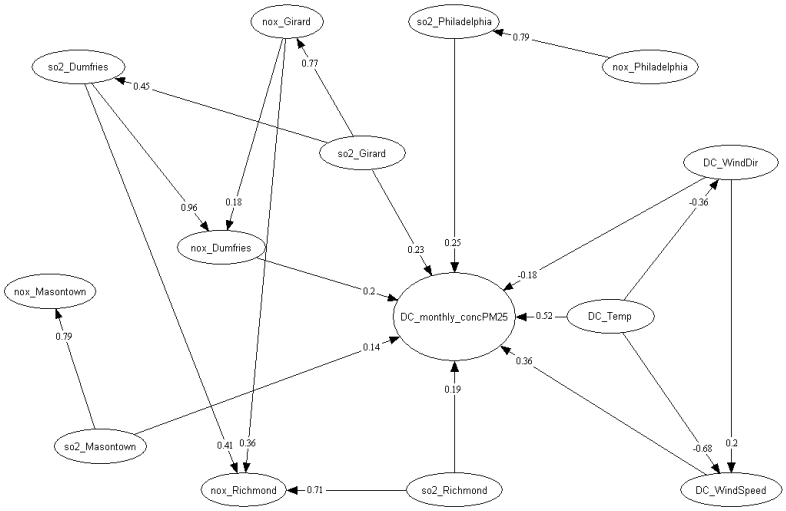


Figure 5.12: *BBN on 14 nodes with 20 arcs.*

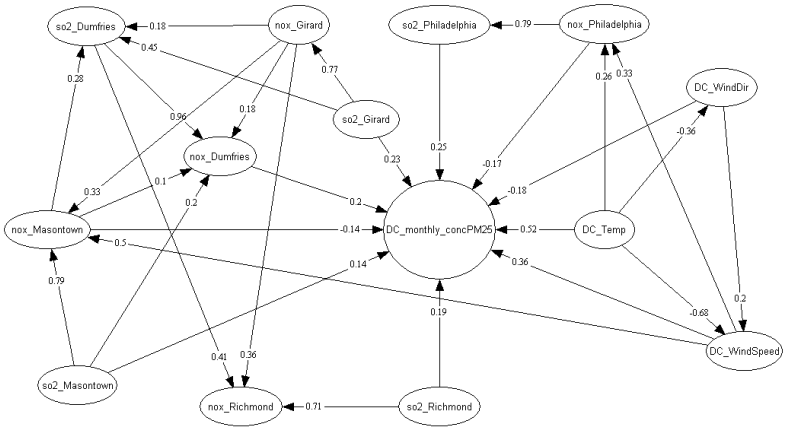


Figure 5.13: *BBN on 14 nodes with 30 arcs.*

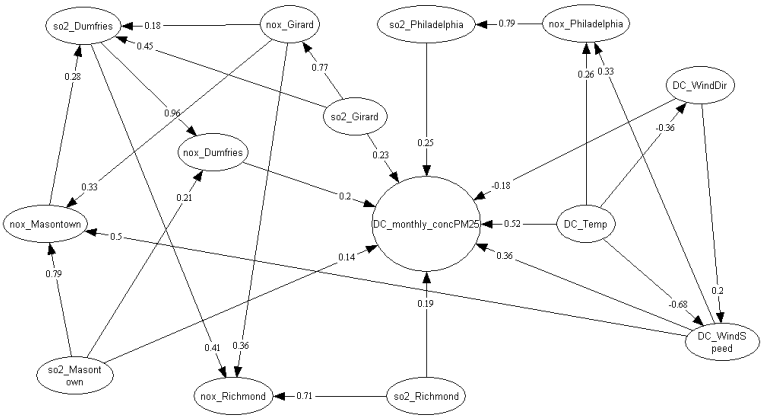


Figure 5.14: *BBN on 14 nodes with 26 arcs.*

very large structures, this operation can be time consuming. In order to avoid this problem we will further present a number of results that will reduce the use of equation 1.2.1. It is known that a BBN induces a (generally non-unique) sampling order and that variable X is independent of variable Y given its parents in the graph, $Pa(Y)$, if X precedes Y in the sampling order. Our aim is to obtain a conditioning set D, which entails conditional independence, smaller than the set of parents. In this case our algorithm to calculate the correlation matrix from partial correlations specified on the BBN will calculate ρ_{XY} from $\rho_{XY;D}$ rather than from $\rho_{XY;Pa(Y)}$.

5.4.1 Notation and Definitions

We begin with the notation used in this section and assume that the reader is familiar with basic concepts of graph theory. The definitions presented in this subsection can be found in the literature, e.g. (Pearl 1988). Capital letters, e.g. X, denote a single variable. Sets of variables are denoted in bold, e.g. $\mathbf{A}$. The sets of ancestors, children and descendants of X are expressed as $An(X)$, $Ch(X)$ and $Desc(X)$, respectively. We consider X an ancestor of itself, i.e.: $An(X) = X \cup \bigcup_{Y \in Pa(X)} An(Y)$. To describe conditional independence between variables X and Y given Z we write $X \perp Y|Z$. $X \perp Y$ if $X \perp Y|\emptyset$. Moreover, $X \not\perp Y|Z$ means that X and Y are not conditionally independent given Z. Hence they are conditionally dependent. $\wp$ denotes an undirected *path*.

A joint distribution represented by a BBN must satisfy a set of independence constraints imposed by the structure of the graph. A graphical criterion that characterises all of these structural independence constraints is the *d-separation* criterion.

Definition 5.4.1. *If* $\boldsymbol{A}$, $\boldsymbol{B}$ *and* $\boldsymbol{C}$ *are three disjoint subsets of nodes in a BBN, then* $\boldsymbol{C}$ *is said to d-separate* $\boldsymbol{A}$ *from* $\boldsymbol{B}$ *if there is no path between a node in* $\boldsymbol{A}$ *and a node in* $\boldsymbol{B}$ *along which the following to conditions hold:*

- *every node with converging arrows is in C or has a descendant in C and*
- *every other node is outside C.*

Figure 5.15 explains the above definition graphically.
If a path satisfies the condition above, it is said to be *active*. Otherwise it is said to be blocked by $\mathbf{C}$. Two variables, X and Y are d-separated if no path between them is active. X and Y are called *d-connected* if there is any active path between them.

In (Spirtes, Glymour, and Scheines 1993) a node with converging arrows is called a *collider*. A *colliderless path* is a path that does not contain any collider.

If X and Y are d-separated by Z we will write $D_{sep}(X;Y|Z)$.

Remark 5.4.1. $D_{sep}(X;Y|\emptyset)$ *implies the absence of a colliderless path between* X *and* Y.

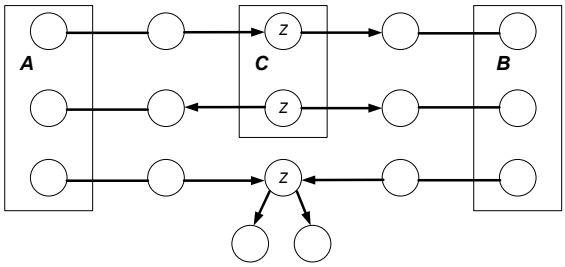


Figure 5.15: *D-separation of* ***A*** *&* ***B*** *by* ***C****.*

5.4.2 Minimal d-separation Set

The d-separation described above provides a very useful connection between a BBN structure and the corresponding set of distributions that can be represented with that structure. In particular, (Pearl 1988) shows that if $D_{sep}(X;Y|Z)$ in a BBN structure, then for any distribution that can be represented by that structure $(X \perp Y|Z)$. Therefore, the absence of an arc guarantees a set of independence facts. On the other hand, the existence of an arc between variables X and Y in the graph, does not guarantee that the BBN will exhibit dependence between X and Y. To ensure this dependence one will have to make the assumption of *faithfulness*. A distribution is *faithful* to a BBN if $X \perp Y|Z$ implies $D_{sep}(X;Y|Z)$. This means that there is a BBN structure, such that the independence relationships among the variables in the distribution are exactly those represented by the BBN by means of the d-separation criterion. We will further present a number of results that can help us to reduce the set of conditioning variables that guarantees conditional independence between X and Y.

Proposition 5.4.1. *Let X and Y be two nodes of a BBN. Then $X \perp Y|An(X) \cap An(Y)$.*

Proof. $X \in An(Y) \Rightarrow X \in An(X) \cap An(Y) \Rightarrow X \perp Y|An(X) \cap An(Y)$. The same argument holds if $Y \in An(X)$.

Let us assume that $X \notin An(Y)$ & $Y \notin An(X)$. The paths between X and Y go through:

1. $An(X) \cap An(Y)$, or
2. $An(X) \cap Desc(Y)$, or
3. $Desc(X) \cap An(Y)$, or
4. $Desc(X) \cap Desc(Y)$.

All paths in the situations 2, 3 and 4 contain a collider. From Definition 5.4.1 it follows that $An(X) \cap An(Y)$ d-separates X from Y, hence $X \perp Y | An(X) \cap An(Y)$. □

The above proposition is not always useful, as the intersection of the ancestors of X and Y may contain more variables than $Pa(Y)$.

Proposition 5.4.2. *Let X and Y be two nodes of a BBN. Under the faithfulness assumption, if $X \perp Y$ then $An(X) \cap An(Y) = \emptyset$.*

Proof. $X \perp Y \Rightarrow D_{sep}(X; Y|\emptyset)$. Remark 5.4.1 implies that each path between X and Y contains a collider. Let us assume $An(X) \cap An(Y) \neq \emptyset$ and let $Z \in An(X) \cap An(Y)$. Then, there exist a path $\wp$ from X to Y, through Z such that, $\wp$ does not contain a collider. This contradiction concludes the proof. □

From the previous two propositions we can conclude that under the faithfulness assumption $X \perp Y$ iff $An(X) \cap An(Y) = \emptyset$.

Proposition 5.4.3. *Let X be a node of a BBN and $Pa(X) = \boldsymbol{A} \cup \boldsymbol{B}$ such that $\boldsymbol{A} \cap \boldsymbol{B} = \emptyset$ and $\boldsymbol{A} \perp \boldsymbol{B}$. Under the faithfulness assumption, if $Y \in An(\boldsymbol{A})$, then $X \perp Y | \boldsymbol{A}$*

Proof. If $Y \in \mathbf{A}$, then $X \perp Y|\mathbf{A}$. Let us consider the case when $Y \notin \mathbf{A}$, then $Y \in An(\mathbf{A}) \setminus \mathbf{A}$.
$\mathbf{A} \perp \mathbf{B} \Rightarrow An(\mathbf{A}) \cap An(\mathbf{B}) = \emptyset \Rightarrow An(\mathbf{A}) \perp An(\mathbf{B})$. Because $\mathbf{B} \subset An(\mathbf{B})$ we conclude that $An(\mathbf{A}) \perp \mathbf{B}$. But $\{\mathbf{A}, Y\} \subset An(\mathbf{A})$. Then $\{\mathbf{A}, Y\} \perp \mathbf{B}$. Using this and the fact that $\mathbf{A} \perp \mathbf{B}$, we can write:

$$P(Y|\mathbf{A},\mathbf{B}) = \frac{P(Y,\mathbf{A},\mathbf{B})}{P(\mathbf{A},\mathbf{B})} = \frac{P(Y,\mathbf{A}|\mathbf{B})\,P(\mathbf{B})}{P(\mathbf{A})\,P(\mathbf{B})} = \frac{P(Y,\mathbf{A})}{P(\mathbf{A})} = P(Y|\mathbf{A}). \quad (5.4.1)$$

This means that $Y \perp \mathbf{B}|\mathbf{A}$. Using also $Y \perp X|(\mathbf{B},\mathbf{A})$, we can conclude $Y \perp (X,\mathbf{B})|\mathbf{A}$ (see (Cowell, Dawid, Lauritzen, and Spiegelhalter 1999; Whittaker 1990)). This implies $X \perp Y|\mathbf{A}$. □

We will further define the *boundary* of the intersection of ancestors of two nodes in a BBN with respect to one of them as follows:

$$bd_X(An(X) \cap An(Y)) = \{Z \in An(X) \cap An(Y) \ : \ \exists\, Ch(Z) \in An(X) \setminus An(Y)\}.$$

Similarly:

$$bd_Y(An(X) \cap An(Y)) = \{Z \in An(X) \cap An(Y) \ : \ \exists\, Ch(Z) \in An(Y) \setminus An(X)\}.$$

The proposition below shows that instead of taking the intersection of ancestor sets of X and Y as in Proposition 5.4.1 it is enough to consider the boundary of this intersection.

Proposition 5.4.4. *Let X and Y be two nodes of a BBN such that $X \notin An(Y)$ and $Y \notin An(X)$. Under the faithfulness assumption:*

$$X \perp Y | bd_X\left(An(X) \cap An(Y)\right) \text{ and}$$
$$X \perp Y | bd_Y\left(An(X) \cap An(Y)\right).$$

Proof. By symmetry, it suffice to prove only one of the relations above. Let $\mathbf{C} = An(X) \cap An(Y)$ and $\mathbf{C}^* = bd_X\left(An(X) \cap An(Y)\right)$. It follows that $\mathbf{C}^* \subseteq \mathbf{C}$ and $X \perp Y | \mathbf{C}$. Let us assume $X \not\perp Y | \mathbf{C}^*$. Then there exist an active path $\wp$ between X and Y. Because $D_{sep}(X; Y | \mathbf{C})$, the path $\wp$ must be blocked by a node $Z \in \mathbf{C} \setminus \mathbf{C}^*$. Then $Z \in An(X) \cap An(Y)$ such that any $Ch(Z)$ belongs either to $An(X) \cap An(Y)$, or to $An(Y) \setminus An(X)$. Since $X \notin (An(X) \cap An(Y))$ it follows that any path going from Z along $\wp$ has to go through a node from $\mathbf{C}^*$ in order to reach X which contradicts the fact that $\wp$ is an active path. It follows that the assumption made is false and $\mathbf{C}^*$ is a separator for X and Y. □

Corollary 5.4.1. *In the context of the previous proposition:*

- *If $X \in An(Y)$ then :*
  - $X \perp Y | bd_Y\left(An(X) \cap An(Y)\right)$
  - $bd_X\left(An(X) \cap An(Y)\right) = \emptyset$
- *If $Y \in An(X)$ then:*
  - $X \perp Y | bd_X\left(An(X) \cap An(Y)\right)$
  - $bd_Y\left(An(X) \cap An(Y)\right) = \emptyset$

Under the faithfulness assumption, the proof of the above corollary is trivial.

If we are in the conditions of Proposition 5.4.3 we definitely have a smaller conditioning set then the set of parents. If, on the other hand, we want to calculate the correlation of two nodes which are non ancestors of one another, we can compare the set of parents with the *boundaries* of the intersection of ancestors and decide which conditioning set will facilitate the calculation.

As an example consider the BBN from Figure 5.2. Choose the following sampling order: so2_Masontown (1), so2_Girard (2), DC_Temp (3), so2_Richmond (4), nox_Girard (5), DC_WindDir (6), DC_WindSpeed (7), nox_Masontown (8), nox_Philadelphia (9), so2_Dumfries (10), so2_Philadelphia (11), nox_Richmond (12), nox_Dumfries (13), DC_monthly_concPM25 (14). Using this sampling order and referring the variables with their indices in the sampling order we can write two relations:

- $14 \perp 12 | Pa(14)$ and
- $14 \perp 12 | bd_{12}\left(An(12) \cap An(14)\right)$.

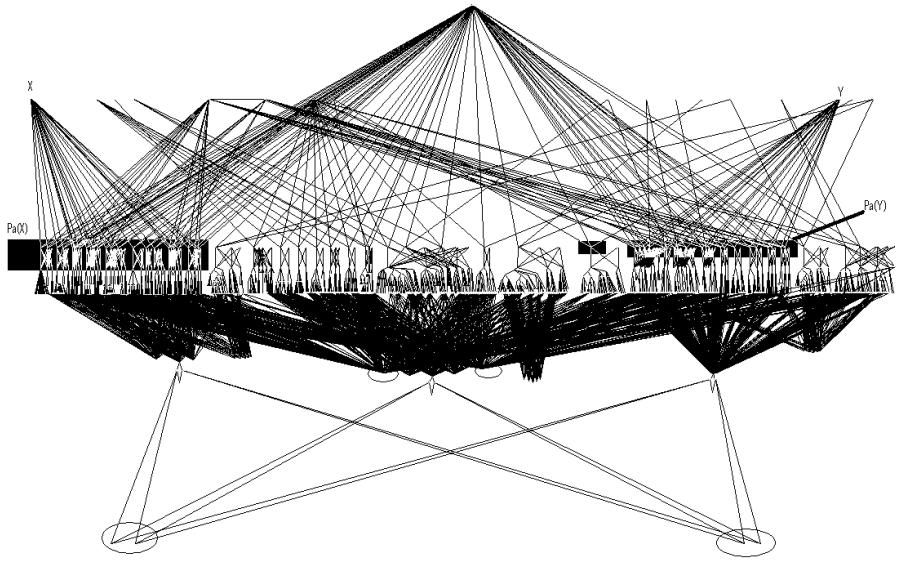


Figure 5.16: *BBN for the CATS model showing the variables X and Y and the sets $Pa(X)$, $Pa(Y)$, and $An(X) \cap An(Y)$.*

The set $Pa(14)$ contains 8 variables, whereas the set $bd_{12}\left(An(12) \cap An(14)\right)$ contains only 3 variables.

Let us now consider the BBN from the *CATS* model, introduced in Chapter 4. We select two nodes from this BBN, X and Y. The nodes together with their parents are showed in Figure 5.16. The sets of parents are marked with black rectangles. The resolution of this picture does not permit a detailed picture of the parents sets, hence the rectangles are just an indication. Let us assume that Y is before X in a sampling order. X has 25 parents[10]. The set $An(X) \cap An(Y)$ consists in only 6 nodes, enclosed in 4 ellipses in Figure 5.16. In this example $bd_X\left(An(X) \cap An(Y)\right) = An(X) \cap An(Y)$.

It is clear that with the above results we can dramatically reduce the use of equation 1.2.1 in calculating the correlation matrix. However we also have to account for the time spent to collect information about the ancestors of each node.

[10]We can also choose a sampling order where X is before Y. Y has 22 parents, therefore the situations are similar.

Chapter 6

Conclusions

6.1 Retrospect

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.

BBNs provide a general methodology for approaching these problems. The graphical (visual) nature of BBNs facilitates the understanding of key linkages between the variables involved in the model.

Until recently, continuous BBNs were restricted to the joint normal distribution. Often times the variables involved in a real life problem are far from normal. Therefore we need an approach that is general, in the sense that it applies to any continuous variables. Kurowicka and Cooke (2004) advanced the copula-vine approach to non-parametric continuous 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 any copula with the zero independence property. The (conditional) rank correlations assigned to the edges are algebraically independent. 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 needs not be revised if the univariate distributions are changed.

This approach is general and allows traceable and defensible quantification methods, but the BBNs must be evaluated by Monte Carlo simulation. Updating such a BBN requires the re-sampling of the whole structure. This motivates the introduction of a hybrid method that samples the continuous BBN once, and then discretizes this so as to enable fast updating. In this way we combine the reduced assessment burden and modelling flexibility of the continuous BBNs with the fast updating algorithms of discrete BBNs. The hybrid method proposed here has

the advantage that we only have to sample once, and any copula with the zero independence property, that realises the entire range of dependence can be used in the sampling procedure. However, sampling large complex structures only once can still involve time consuming numerical calculations. Hence a new sampling protocol based on normal vines is developed. The normal copula is used to realise the dependence structure specified via (conditional) rank correlations on the continuous BBN.

The normal copula vine method can be used to sample any non-parametric continuous BBN and stipulate its joint distribution in a fast and flexible way. A very attractive feature of this method is that conditioning (updating) can be done analytically.

We have extended this approach to include ordinal discrete random variables which can be written as monotone transforms of uniform variables. 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 we first study the relationship between these two ranks. We formulate a generalisation of the population version of Spearman's rank correlation for the case of ordinal discrete random variables. A class of bivariate discrete distributions can be constructed by specifying the marginal distributions and a copula. One parameter copulae can be parameterised by their rank correlation. We have established the relation between the rank correlation of the discrete variables and the rank correlation of the chosen copula, with more emphasis on the normal copula.

In situations when data does not exist or is very sparse we must rely on expert judgement to define the graphical structure and assess required parameters. However, if data is available we would like to extract a *good* fitting model from the data. The second part of this thesis treats BBNs as tools for mining ordinal multivariate data. There already is a large body of scientific work available on this subject. Nevertheless, to our knowledge, the few methods that can handle non-parametric continuous variables (e.g., Margaritis 2005) can hardly be applied in domains with a large number of variables that are densely connected. Moreover the existing BBN structure learning algorithms are slow, both in theory (Chickering et al. 1994) and in practice e.g., most constraint-based algorithms require an exponential number of conditional independence tests.

We have proposed a method that overcomes these problems; it can handle a large number of continuous variables, without making any assumption about their marginal distributions, in a very fast manner. Once we have learned the BBN from data, we can further use it for prediction, or diagnosis by employing the methods described in the first part of this thesis.

6.2 Prospect

Even though the conditional rank correlation representation of influence, with the joint normal copula, can handle large problems without making any assumptions about the univariate distributions, without excessive assessment and computational burden there are still many interesting open issues on non-parametric BBNs. The requisite computational speed can only be obtained with the joint normal copula. It is hoped that other copulae could be used in the future. If so, it would be interesting to study how sensitive the relationship between the rank correlation of two discrete variables and the rank correlation of their underlying uniforms, is to the choice of the copula (or marginal distributions). What would this relation become in case of a not ordered copula?

It is obvious that the class of bivariate discrete distributions constructed by specifying the marginals and a copula depends on the properties of the chosen copula. It would be valuable to determine what kind of restrictions we introduce using this construction.

As in the case of the rank correlation between two continuous variables, a value for the rank correlation between two discrete variables can be either obtained from data, or form experts. The technique for eliciting (conditional) rank correlations for discrete variables is still an open issue.

Another challenge is to perform conditionalisation on functional nodes. In the current framework, such conditionalisation can only be preformed on samples generated by the BBN. More sophisticated sampling and computational methods might offer other possibilities.

From the inference point of view, learning the structure of a BBN from data requires a suitable measure of multivariate dependence. Our choice is the determinant of the correlation matrix. As mentioned and motivated previously, we actually work with the determinant of the rank correlation matrix. This measure is not such an intuitive one. Maybe a better measure of multivariate dependence would be the mutual information, but calculating the empirical mutual information for large dimensions is a complicated task.

Only lightly touched upon in this thesis are the statistical tests for the two validation steps of our data mining approach. More reliable statistical tests would be of great interest.

Chapter 7

Appendix

7.1 UniNet

UniNet is a standalone uncertainty analysis software package that was developed to support the *CATS* project. The software will be shortly available free from http://dutiosc.twi.tudelft.nl/∼ risk/, together with supporting scientific documentation. Its main focus is dependence modelling for high dimensional distributions. Random variables can be coupled using a BBN. Following is a basic walkthrough for creating and working with a small BBN.

UniNet is launched from *Start/All Programs/Uninet/Uninet*. There are two main views in UniNet: the *Random Variable View* and the *Bayesian Belief Net View*. UniNet starts up in the *Random Variable View*[1] and it allows the user to choose the variables needed for the model. The names of the variables are by default Vi, but they can be changed by double-clicking the name in the list of variables. The distribution of the variables can also be chosen, with corresponding parameters (Figure 7.1).

UniNet can handle continuous distributions, discrete distributions (see Figure 7.2), and distributions imported from a sample file.

Once the random variables are created, they can be used to form a BBN. Assume 4 random variables are created (the last of which discrete) and the *Bayesian Belief Net View* is selected from the *View* menu. On the right hand side of this view there is a list containing the random variables. This list can be shown or hidden[2]. The four variables can be dragged onto the blank canvas (Figure 7.3).

It can be noticed that the name of the V4 node is in italics, this is to indicate that it is a discrete BBN node. The node name[3] and/or description can be modified by the user. The PDF and CDF histograms of the node can be viewed.

When the *Arc* button is pressed UniNet enters the *Add arc* edit mode. Clicking on V4 and then on V2, V4 will become a parent of V2. The arc appears, and

[1] When loading a model it will switch to the *Bayesian Belief Net View*.

[2] It is not much use when it is empty, for instance.

[3] Node names must be unique and can only include letters, digits and underscore characters.

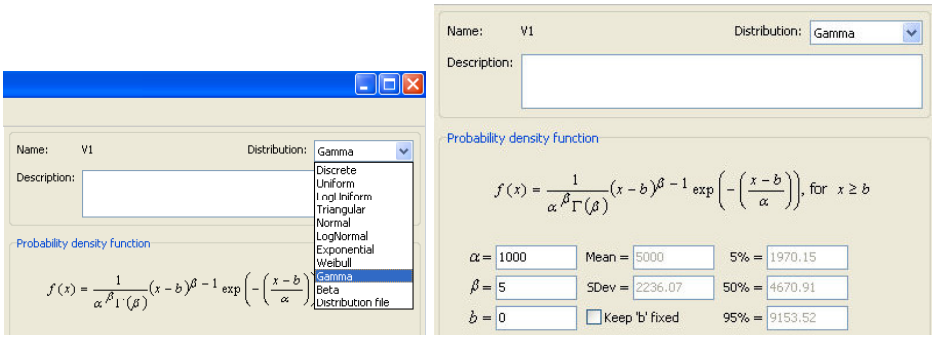


Figure 7.1: *Choosing a random variable's distribution.*

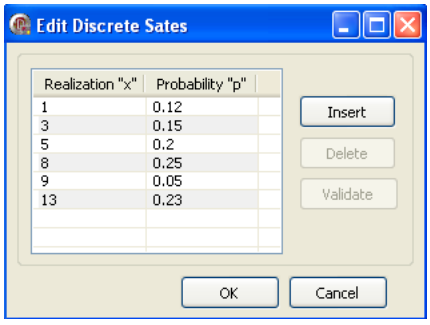


Figure 7.2: *Creating a discrete random variable.*

the zero displayed on it is the rank correlation coefficient associated with the arc (see Figure 7.4).

The rank correlation coefficient associated with an arc can be set by right-clicking on the child node and selecting *Dependence Info*. The rank correlation between V4 and V2 and the rank correlation between V4 and V3 are unconditional rank correlations. In the dependence info for node V1 only the first rank correlation is unconditional, while the following two are conditional, with the conditioning set consisting of the previous parents in the parent ordering (the parent ordering is from top to bottom in the list; this can be changed, as we will see further). If some (non-zero) values are set for the three rank correlations and then the order of the last 2 parents is changed, the bottom two correlation coefficients are reset to zero, and a message is displayed in the lower part of the window explaining why they were reset (Figure 7.5).

When the BBN is completely specified, the correlation matrix can be calculated via *View/Compute Correlation Matrix*. Figure 7.6 shows the rank correla-

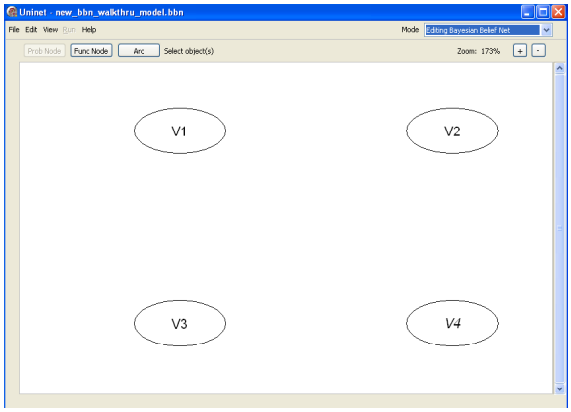


Figure 7.3: *Adding nodes to the canvas.*

tion matrix of the random variables corresponding to all the nodes in the BBN. The user can reorder the rows/columns (useful when the BBN grows) and export the matrix to a text file. Another useful feature is the possibility to highlight parent-child correlations (cells with orange background colour) or highlight the k highest non parent-child correlations (cells with blue background colour)[4].

This correlation matrix is computed from the graph structure and (conditional) rank correlations associated with the arcs using the normal copula vine approach. The matrix is of course necessary for sampling the BBN (conditionally or not) and, once it is calculated, UNINET re-computes it only if the BBN structure changes.

The ellipse shape is not the only one that BBN nodes can have. They can also be viewed as histograms. Figure 7.7 shows the histogram, mean and standard deviation of each variable (node).

From the *Mode* drop down combo box in the top-right of the window, one can select *Analytical Conditioning*. So far, UNINET has been in the operating mode *Editing Bayesian Belief Net*. In this mode, the user can create/modify the BBN and the (conditional) correlation coefficients associated with the arcs.

When switching to *Analytical Conditioning*[5] operating mode it can be noticed that a quick sampling action takes place. That is because, in the *Analytical Conditioning* operating mode, an underlying sample is always in existence. In this operating mode, the user can set one or more of the nodes to point values within their range, and then propagate this evidence through the graph (in the case of discrete nodes, the node can be set to one of its states).

[4]Both options are available in the *Correlation Matrix* window in full view mode - click text *More >>* if not visible.

[5]The conditioning is done using the normal copula vine approach, hence in a analytical way.

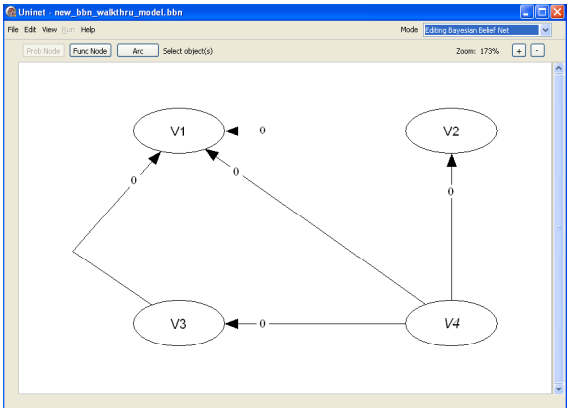


Figure 7.4: *BBN with arcs.*

Upon right-clicking on node V1 and selecting *Conditionalise*, the histogram for the PDF of this continuous node is displayed. The two values at the ends of the horizontal axis are the lower and upper bounds for the range of this variable. The value on which we wish to conditionalise is 9000. The node has now a constant point value, and it has changed into a gray rounded rectangle displaying this value. When the *Update* button is pressed, a new sample is created. The marginal distributions of the other three nodes, V2, V3 and V4, have changed (Figure 7.8). The gray histograms in the background describe the unconditional distributions of each node, while the black histograms in the foreground describe the conditional distributions for each node. The background, unconditional histograms are referred to as *shadow histograms* in UniNet. The mean and standard deviation of each node have changed too.

Apart from probabilistic nodes, a BBN can contain so-called *functional nodes* as well. These are nodes which are functions of their parents. In order to create a functional node, the user needs to switch the operating mode of UniNet back to *Editing Bayesian Belief Net*, press *Func Node* button at the top of UniNet 's window, and click somewhere on the area of the BBN canvas. A functional node will be created and it will be given a default name starting with FN_. The input variables for node FN_1 are declared by creating an arc originating from a node and ending at FN_1 (see Figure 7.9). A functional node cannot be a parent of a probabilistic node. UniNet will not allow for such a construction.

The functional relationship represented by this node can be edited in the *Dependence Info* window available for this node when right-clicked. Tab *Nodes* lists all random variables available for use in the formula. Tabs *Functions*, *Operators*, *Constants* contain the list of available terms for building a formula. After a formula is assigned to the functional node, UniNet is ready to sample the BBN

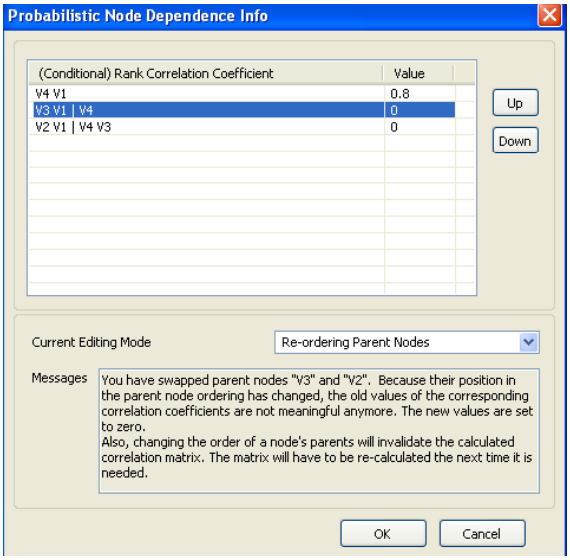


Figure 7.5: *Dependence Information: Re-ordering parent nodes.*

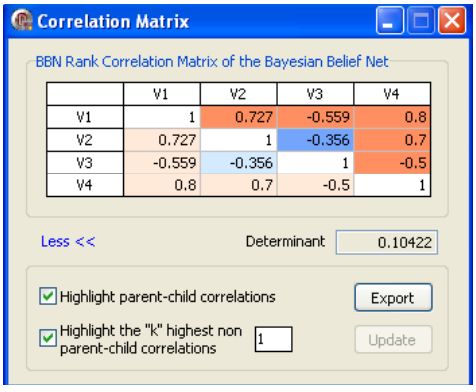


Figure 7.6: *The rank correlation matrix for the entire joint distribution.*

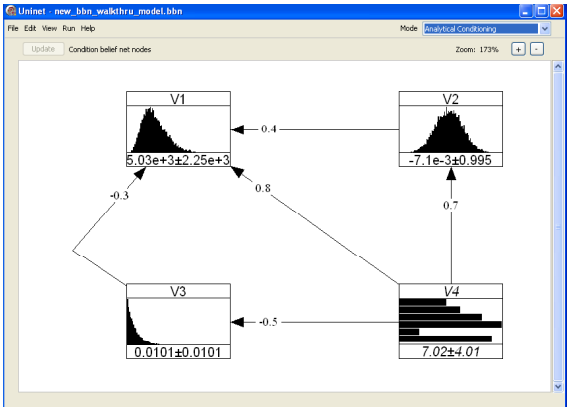


Figure 7.7: *Viewing nodes as histograms.*

structure.

The way to data mine a sample file will be very briefly described here, since its main features are presented in Chapter 5.

A text file with the extension **.sae* should be available. This is a simple, comma-separated sample file, with the names of the random variables on the first row, and multidimensional samples on the following rows. Upon selecting menu *File/New/Data Mining File* and choosing the **.sae* file the sample file is loaded. The current view becomes the *Bayesian Belief Net view*. The random variables are imported and can be used now for creating a BBN. On the right hand side of this view there is a list containing the random variables. This list can be shown or hidden from the menu *View/Random Variables*. Drag and drop the random variables onto the blank canvas. After arcs are also added to the BBN, UNINET calculates the (conditional) rank correlations associated to the arcs, from the sample file. When the BBN is completed all functionality presented earlier in this section can be applied.

7.2 PROOF OF THEOREM 3.4.1

Proof of Theorem 3.4.1: For simplicity, all calculations are done for the case n=m. In order to prove the expression of $P_c - P_d$ from the theorem, we first acquire an intermediate result:

$$P_c - P_d = \sum_{i,j=1}^{m-1} p_{ij} \left(p_{i+} + 2 \sum_{k=i+1}^{m-1} p_{k+} + p_{m+} \right) \left(p_{+j} + 2 \sum_{l=j+1}^{m-1} p_{+l} + p_{+m} \right) - \sum_{i,j=1}^{m-1} p_{i+} p_{+j}. \tag{7.2.1}$$

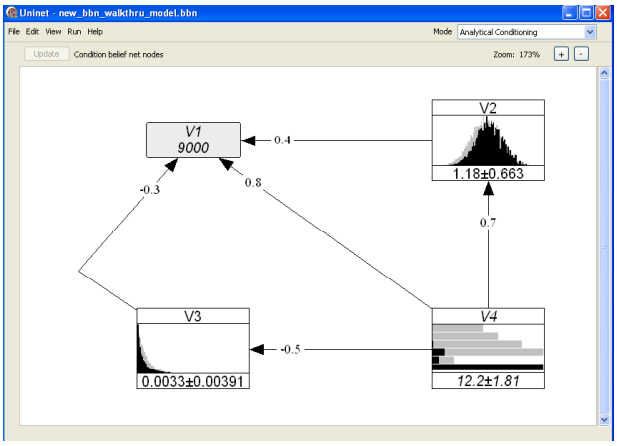


Figure 7.8: *Conditionalising on one node.*

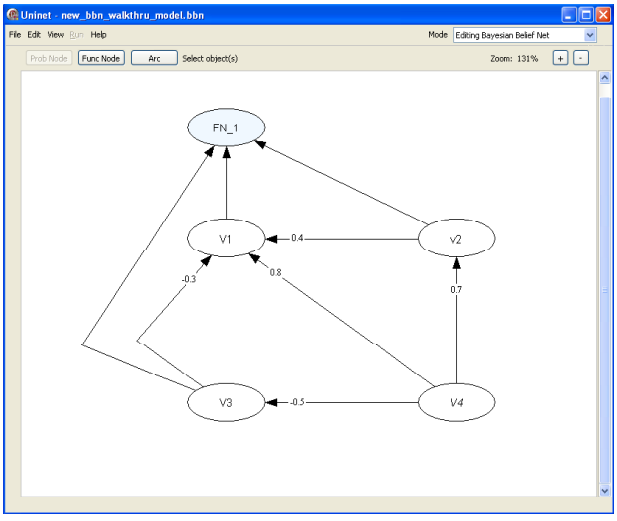


Figure 7.9: *Creating a functional node.*

Proof of equation (7.2.1): We start from equation (3.3.1) and rewrite the sums for the indexes running from 1 to $m-1$ (rather than to m):

$$
\begin{aligned}
P_c \;-\; P_d &= \sum_{i,j=1}^{m} p_{ij}\Big(\sum_{k\neq i}\sum_{l\neq j} sign\,((k-i)(l-j))\,q_{kl}\Big)\\
&= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k\neq i}\sum_{l\neq j} sign\,((k-i)(l-j))\,q_{kl}\Big)\\
&+ \sum_{i=1}^{m-1}\Big(p_{i+}-\sum_{j=1}^{m-1}p_{ij}\Big)\Big(\sum_{k\neq i}\sum_{l\neq m} sign\,((k-i)(l-m))\,q_{kl}\Big)\\
&+ \sum_{j=1}^{m-1}\Big(p_{+j}-\sum_{i=1}^{m-1}p_{ij}\Big)\Big(\sum_{k\neq m}\sum_{l\neq j} sign\,((k-m)(l-j))\,q_{kl}\Big)\\
&+ \Big(p_{m+}+p_{+m}-1+\sum_{i,j=1}^{m-1}p_{ij}\Big)\sum_{k,l=1}^{m-1} q_{kl}\\
&= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k\neq i}\sum_{l\neq j} sign\,((k-i)(l-j))\,q_{kl}\Big)\\
&+ \sum_{i=1}^{m-1}\Big(p_{i+}-\sum_{j=1}^{m-1}p_{ij}\Big)\Big(\sum_{k\neq i}\sum_{l=1}^{m-1} sign(i-k)q_{kl}\Big)\\
&+ \sum_{j=1}^{m-1}\Big(p_{+j}-\sum_{i=1}^{m-1}p_{ij}\Big)\Big(\sum_{k=1}^{m-1}\sum_{l\neq j} sign(j-l)q_{kl}\Big)\\
&+ \Big(p_{m+}+p_{+m}-1+\sum_{i,j=1}^{m-1}p_{ij}\Big)\sum_{k,l=1}^{m-1} q_{kl}\\
&= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{j-1} q_{kl}-\sum_{k=1}^{i-1}\sum_{l=j+1}^{m} q_{kl}-\sum_{k=i+1}^{m}\sum_{l=1}^{j-1} q_{kl}+\sum_{k=i+1}^{m}\sum_{l=j+1}^{m} q_{kl}\Big)\\
&+ \sum_{i=1}^{m-1} p_{i+}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{m-1} q_{kl}-\sum_{k=i+1}^{m}\sum_{l=1}^{m-1} q_{kl}\Big)-\sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{m-1} q_{kl}-\sum_{k=i+1}^{m}\sum_{l=1}^{m-1} q_{kl}\Big)\\
&+ \sum_{j=1}^{m-1} p_{+j}\Big(\sum_{k=1}^{m-1}\sum_{l=1}^{j-1} q_{kl}-\sum_{k=1}^{m-1}\sum_{l=j+1}^{m} q_{kl}\Big)-\sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{m-1}\sum_{l=1}^{j-1} q_{kl}-\sum_{k=1}^{m-1}\sum_{l=j+1}^{m} q_{kl}\Big)\\
&+ p_{m+}\sum_{k,l=1}^{m-1} q_{kl}+p_{+m}\sum_{k,l=1}^{m-1} q_{kl}-\sum_{k,l=1}^{m-1} q_{kl}+\sum_{i,j=1}^{m-1} p_{ij}\sum_{k,l=1}^{m-1} q_{kl}.
\end{aligned}
$$

We will further collect all terms which contain p_{ij}'s, and use the relation $q_{ij}{=}p_{i+}p_{+j}$ to rewrite some of the terms in the expression above:

$$
\begin{aligned}
P_c - P_d &= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{j-1} q_{kl} - \sum_{k=1}^{i-1}\sum_{l=j+1}^{m} q_{kl} - \sum_{k=i+1}^{m}\sum_{l=1}^{j-1} q_{kl} \\
&+ \sum_{k=i+1}^{m}\sum_{l=j+1}^{m} q_{kl} - \sum_{k=1}^{i-1}\sum_{l=1}^{m-1} q_{kl} + \sum_{k=i+1}^{m}\sum_{l=1}^{m-1} q_{kl} - \sum_{k=1}^{m-1}\sum_{l=1}^{j-1} q_{kl} \\
&+ \sum_{k=1}^{m-1}\sum_{l=j+1}^{m} q_{kl} + \sum_{k,l=1}^{m-1} q_{kl}\Big) - \sum_{k,l=1}^{m-1} q_{kl} + \sum_{i=2}^{m-1}\sum_{k=1}^{i-1}\sum_{l=1}^{m-1} p_{i+}p_{k+}p_{+l} \\
&- \sum_{i=1}^{m-1}\sum_{k=i+1}^{m}\sum_{l=1}^{m-1} p_{i+}p_{k+}p_{+l} + \sum_{j=2}^{m-1}\sum_{k=1}^{m-1}\sum_{l=1}^{j-1} p_{+j}p_{k+}p_{+l} - \sum_{j=1}^{m-1}\sum_{k=1}^{m-1}\sum_{l=j+1}^{m} p_{+j}p_{k+}p_{+l} \\
&+ \sum_{k=1}^{m-1}\sum_{l=1}^{m-1} p_{m+}p_{k+}p_{+l} + \sum_{k=1}^{m-1}\sum_{l=1}^{m-1} p_{+m}p_{k+}p_{+l} \\
&= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{j-1} q_{kl} - \sum_{k=1}^{i-1}\sum_{l=j+1}^{m} q_{kl} - \sum_{k=i+1}^{m}\sum_{l=1}^{j-1} q_{kl} + \sum_{k=i+1}^{m}\sum_{l=j+1}^{m} q_{kl} \\
&- \sum_{k=1}^{i-1}\sum_{l=1}^{m-1} q_{kl} + \sum_{k=i+1}^{m}\sum_{l=1}^{m-1} q_{kl} - \sum_{k=1}^{m-1}\sum_{l=1}^{j-1} q_{kl} + \sum_{k=1}^{m-1}\sum_{l=j+1}^{m} q_{kl} + \sum_{k,l=1}^{m-1} q_{kl}\Big) \\
&- \sum_{i,j=1}^{m-1} q_{ij} + \sum_{i=2}^{m-1}\sum_{k=1}^{i-1}\sum_{l=1}^{m-1} p_{i+}p_{k+}p_{+l} - \sum_{i=1}^{m-2}\sum_{k=i+1}^{m-1}\sum_{l=1}^{m-1} p_{i+}p_{k+}p_{+l} \\
&+ \sum_{i=1}^{m-1}\sum_{j=2}^{m-1}\sum_{l=1}^{j-1} p_{i+}p_{+j}p_{+l} - \sum_{i=1}^{m-1}\sum_{j=1}^{m-2}\sum_{l=j+1}^{m-1} p_{i+}p_{+j}p_{+l} - \sum_{i=1}^{m-1}\sum_{l=1}^{m-1} p_{i+}p_{m+}p_{+l} \\
&- \sum_{j=1}^{m-1}\sum_{k=1}^{m-1} p_{+j}p_{k+}p_{+m} + \sum_{k=1}^{m-1}\sum_{l=1}^{m-1} p_{k+}p_{m+}p_{+l} + \sum_{k=1}^{m-1}\sum_{l=1}^{m-1} p_{+l}p_{k+}p_{+m}.
\end{aligned}
$$

Rearranging the terms from the triple sums will result in their cancelation. We will manipulate the terms which are multiplied by p_{ij}'s.

$$
\begin{aligned}
P_c - P_d &= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{j-1} q_{kl} - \sum_{k=1}^{i-1}\sum_{l=j+1}^{m} q_{kl} - \sum_{k=i+1}^{m}\sum_{l=1}^{j-1} q_{kl} + \sum_{k=i+1}^{m}\sum_{l=j+1}^{m} q_{kl} \\
&- \sum_{k=1}^{i-1}\sum_{l=1}^{m-1} q_{kl} + \sum_{k=i+1}^{m}\sum_{l=1}^{m-1} q_{kl} - \sum_{k=1}^{m-1}\sum_{l=1}^{j-1} q_{kl} + \sum_{k=1}^{m-1}\sum_{l=j+1}^{m} q_{kl} + \sum_{k=1}^{i-1}\sum_{l=1}^{m-1} q_{kl} \\
&+ \sum_{l=1}^{m-1} q_{il} + \sum_{k=i+1}^{m}\sum_{l=1}^{j-1} q_{kl} - \sum_{l=1}^{j-1} q_{ml} + \sum_{k=i+1}^{m-1} q_{kj} + \sum_{k=i+1}^{m-1}\sum_{l=j+1}^{m-1} q_{kl}\Big) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j} \\
&= \sum_{i,j=1}^{m-1} p_{ij}\Big(\sum_{k=1}^{i-1}\sum_{l=1}^{j-1} q_{kl} - \sum_{k=1}^{i-1}\sum_{l=j+1}^{m} q_{kl} + 2\sum_{k=i+1}^{m-1}\sum_{l=j+1}^{m-1} q_{kl} + \sum_{k=i+1}^{m-1} q_{km} \\
&+ \sum_{l=j+1}^{m-1} q_{ml} + q_{mm} + \sum_{k=i+1}^{m-1}\sum_{l=1}^{j-1} q_{kl} + \sum_{k=i+1}^{m-1} q_{kj} + \sum_{l=1}^{j-1} q_{ml} + q_{mj} \\
&+ \sum_{k=i+1}^{m-1}\sum_{l=j+1}^{m-1} q_{kl} + \sum_{l=j+1}^{m-1} q_{ml} - \sum_{k=1}^{i-1}\sum_{l=1}^{j-1} q_{kl} - \sum_{k=i+1}^{m-1}\sum_{l=1}^{j-1} q_{kl} - \sum_{l=1}^{j-1} q_{il} \\
&+ \sum_{k=1}^{i-1}\sum_{l=j+1}^{m} q_{kl} + \sum_{k=i+1}^{m-1}\sum_{l=j+1}^{m-1} q_{kl} + \sum_{k=i+1}^{m-1} q_{km} + \sum_{l=j+1}^{m-1} q_{il} + q_{im} + \sum_{l=1}^{j-1} q_{il} + q_{ij} \\
&+ \sum_{l=j+1}^{m-1} q_{il} - \sum_{l=1}^{j-1} q_{ml} + \sum_{k=i+1}^{m-1} q_{kj}\Big) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j} \\
&= \sum_{i,j=1}^{m-1} p_{ij}\big(4\sum_{k=i+1}^{m-1}\sum_{l=j+1}^{m-1} p_{k+}p_{+l} + 2\sum_{k=i+1}^{m-1} p_{k+}p_{+m} + 2\sum_{l=j+1}^{m-1} p_{m+}p_{+l} + p_{m+}p_{+m} \\
&+ 2\sum_{k=i+1}^{m-1} p_{k+}p_{+j} + p_{m+}p_{+j} + 2\sum_{l=j+1}^{m-1} p_{i+}p_{+l} + p_{i+}p_{+m} + p_{i+}p_{+j}\big) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j} \\
&= \sum_{i,j=1}^{m-1} p_{ij}(p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j}.
\end{aligned}
$$

Now we can use the expression for p_{ij} from (3.4.1) in equation (7.2.1):

$$
\begin{aligned}
P_c - P_d &= \sum_{i,j=1}^{m-1}\Big[C_r\Big(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\Big) + C_r\Big(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\Big) - C_r\Big(\sum_{k=1}^{i-1} p_{k+}, \sum_{l=1}^{j} p_{+l}\Big) \\
&- C_r\Big(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j-1} p_{+l}\Big)\Big](p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m})
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{i,j=1}^{m-1} p_{i+}p_{+j} \\
&= \sum_{i,j=1}^{m-2} C_r\Big(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{j} p_{+l}\Big)\Big[(p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}) \\
&+ (p_{(i+1)+} + 2\sum_{k=i+2}^{m-1} p_{k+} + p_{m+})(p_{+(j+1)} + 2\sum_{l=j+2}^{m-1} p_{+l} + p_{+m}) \\
&- (p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+(j+1)} + 2\sum_{l=j+2}^{m-1} p_{+l} + p_{+m}) \\
&- (p_{(i+1)+} + 2\sum_{k=i+2}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m})\Big] \\
&+ \sum_{i=1}^{m-2} C_r\Big(\sum_{k=1}^{i} p_{k+}, \sum_{l=1}^{m-1} p_{+l}\Big)\cdot\Big[(p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+(m-1)} + p_{+m}) \\
&- (p_{(i+1)+} + 2\sum_{k=i+2}^{m-1} p_{k+} + p_{m+})(p_{+(m-1)} + p_{+m})\Big] \\
&+ \sum_{j=1}^{m-2} C_r\Big(\sum_{k=1}^{m-1} p_{k+}, \sum_{l=1}^{j} p_{+l}\Big)\cdot\Big[(p_{(m-1)+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}) \\
&- (p_{(m-1)+} + p_{m+})(p_{+(j+1)} + 2\sum_{l=j+2}^{m-1} p_{+l} + p_{+m})\Big] \\
&+ C_r\Big(\sum_{k=1}^{m-1} p_{k+}, \sum_{l=1}^{m-1} p_{+l}\Big)(p_{(m-1)+} + p_{m+})(p_{+(m-1)} + p_{+m}) - \sum_{i,j=1}^{m-1} p_{i+}p_{+j}.
\end{aligned}
$$

We will use the following notations:

$$
\begin{aligned}
c_1 &= (p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}) \\
&+ (p_{(i+1)+} + 2\sum_{k=i+2}^{m-1} p_{k+} + p_{m+})(p_{+(j+1)} + 2\sum_{l=j+2}^{m-1} p_{+l} + p_{+m}) \\
&- (p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+(j+1)} + 2\sum_{l=j+2}^{m-1} p_{+l} + p_{+m}) \\
&- (p_{(i+1)+} + 2\sum_{k=i+2}^{m-1} p_{k+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m});
\end{aligned}
$$

$$\begin{aligned} c_2 &= (p_{i+} + 2\sum_{k=i+1}^{m-1} p_{k+} + p_{m+})(p_{+(m-1)} + p_{+m}) \\ &- (p_{(i+1)+} + 2\sum_{k=i+2}^{m-1} p_{k+} + p_{m+})(p_{+(m-1)} + p_{+m}); \end{aligned}$$

and

$$\begin{aligned} c_3 &= (p_{(m-1)+} + p_{m+})(p_{+j} + 2\sum_{l=j+1}^{m-1} p_{+l} + p_{+m}) \\ &- (p_{(m-1)+} + p_{m+})(p_{+(j+1)} + 2\sum_{l=j+2}^{m-1} p_{+l} + p_{+m}). \end{aligned}$$

We will rewrite c_1, c_2 and c_3 as follows:

$$\begin{aligned} c_1 &= q_{ij} + 2\sum_{k=i+1}^{m-1} q_{kj} + q_{mj} + 2\sum_{l=j+1}^{m-1} q_{il} + 4\sum_{k=i+1}^{m-1}\sum_{l=j+1}^{m-1} q_{kl} + 2\sum_{l=j+1}^{m-1} q_{ml} + q_{im} \\ &+ 2\sum_{k=i+1}^{m-1} q_{km} + q_{mm} + q_{(i+1)(j+1)} + 2\sum_{k=i+2}^{m-1} q_{k(j+1)} + q_{m(j+1)} + 2\sum_{l=j+2}^{m-1} q_{(i+1)l} \\ &+ 4\sum_{k=i+2}^{m-1}\sum_{l=j+2}^{m-1} q_{kl} + 2\sum_{l=j+2}^{m-1} q_{ml} + q_{(i+1)m} + 2\sum_{k=i+2}^{m-1} q_{km} + q_{mm} - q_{i(j+1)} \\ &- 2\sum_{k=i+1}^{m-1} q_{k(j+1)} - q_{m(j+1)} - 2\sum_{l=j+2}^{m-1} q_{il} - 4\sum_{k=i+1}^{m-1}\sum_{l=j+2}^{m-1} q_{kj} - 2\sum_{l=j+2}^{m-1} q_{ml} - q_{im} \\ &- 2\sum_{k=i+1}^{m-1} q_{km} - q_{mm} - q_{(i+1)j} - 2\sum_{k=i+2}^{m-1} q_{kj} - q_{mj} - 2\sum_{l=j+1}^{m-1} q_{(i+1)l} \\ &- 4\sum_{k=i+2}^{m-1}\sum_{l=j+1}^{m-1} q_{kl} - 2\sum_{l=j+1}^{m-1} q_{ml} - q_{(i+1)m} - 2\sum_{k=i+2}^{m-1} q_{km} - q_{mm} \\ &= q_{ij} + q_{(i+1)(j+1)} + 2q_{(i+1)j} + 2q_{i(j+1)} + 2q_{m(j+1)} - 2q_{(i+1)(j+1)} - 2q_{(i+1)(j+1)} \\ &- 2q_{m(j+1)} - q_{(i+1)j} - q_{i(j+1)} + 4\sum_{k=i+1}^{m-1} q_{k(j+1)} - 4\sum_{k=i+2}^{m-1} q_{k(j+1)} \\ &= (p_{i+} + p_{(i+1)+})(p_{+j} + p_{+(j+1)}); \end{aligned}$$

$$c_2 = q_{i(m-1)} + 2\sum_{k=i+2}^{m-1} q_{k(m-1)} + 2q_{(i+1)(m-1)} + q_{m(m-1)} + q_{im} + 2q_{(i+1)m}$$

$$
\begin{aligned}
&+ \quad 2\sum_{k=i+2}^{m-1} q_{km} + q_{mm} - q_{(i+1)(m-1)} - 2\sum_{k=i+2}^{m-1} q_{k(m-1)} - q_{m(m-1)} - q_{(i+1)m} \\
&- \quad 2\sum_{k=i+2}^{m-1} q_{km} - q_{mm} = (p_{i_+} + p_{(i+1)_+})(p_{+m} + p_{+(m-1)});
\end{aligned}
$$

$$
\begin{aligned}
c_3 \quad &= \quad q_{(m-1)j} + 2\sum_{l=j+2}^{m-1} q_{(m-1)l} + 2q_{(m-1)(j+1)} + q_{(m-1)m} + q_{mj} + 2\sum_{l=j+2}^{m-1} q_{ml} \\
&+ \quad 2q_{m(j+1)} + q_{mm} - q_{(m-1)(j+1)} - 2\sum_{l=j+2}^{m-1} q_{(m-1)l} - q_{(m-1)m} - q_{m(j+1)} \\
&- \quad 2\sum_{l=j+2}^{m-1} q_{ml} - q_{mm} = q_{(m-1)j} + q_{(m-1)(j+1)} + q_{mj} + q_{m(j+1)} \\
&= \quad (p_{(m-1)_+} + p_{m_+})(p_{+j} + p_{+(j+1)}).
\end{aligned}
$$

Hence:

$$
\begin{aligned}
P_c \quad &- \quad P_d = \sum_{i,j=1}^{m-2} C_r\Big(\sum_{k=1}^{i} p_{k_+}, \sum_{l=1}^{j} p_{+l}\Big) \cdot (p_{i_+} + p_{(i+1)_+})(p_{+j} + p_{+(j+1)}) \\
&+ \quad \sum_{i=1}^{m-2} C_r\Big(\sum_{k=1}^{i} p_{k_+}, \sum_{l=1}^{m-1} p_{+l}\Big) \cdot (p_{i_+} + p_{(i+1)_+})(p_{+m} + p_{+(m-1)}) \\
&+ \quad \sum_{j=1}^{m-2} C_r\Big(\sum_{k=1}^{m-1} p_{k_+}, \sum_{l=1}^{j} p_{+l}\Big) \cdot (p_{(m-1)_+} + p_{m_+})(p_{+j} + p_{+(j+1)}) \\
&+ \quad C_r\Big(\sum_{k=1}^{m-1} p_{k_+}, \sum_{l=1}^{m-1} p_{+l}\Big)(p_{(m-1)_+} + p_{m_+})(p_{+(m-1)} + p_{+m}) - \sum_{i,j=1}^{m-1} p_{i_+} p_{+j} \\
&= \quad \sum_{i,j=1}^{m-1} C_r\Big(\sum_{k=1}^{i} p_{k_+}, \sum_{l=1}^{j} p_{+l}\Big) \cdot (p_{i_+} + p_{(i+1)_+})(p_{+j} + p_{+(j+1)}) - \sum_{i,j=1}^{m-1} p_{i_+} p_{+j} \\
&= \quad \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} p_{i_+} p_{+j} \Bigg(C_r\Big(\sum_{k=1}^{i} p_{k_+}, \sum_{l=1}^{j} p_{+l}\Big) + C_r\Big(\sum_{k=1}^{i-1} p_{k_+}, \sum_{l=1}^{j} p_{+l}\Big) \\
&+ \quad C_r\Big(\sum_{k=1}^{i} p_{k_+}, \sum_{l=1}^{j-1} p_{+l}\Big) + C_r\Big(\sum_{k=1}^{i-1} p_{k_+}, \sum_{l=1}^{j-1} p_{+l}\Big) - 1\Bigg) \\
&= \quad \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} p_{i_+} p_{+j} \big(\widetilde{C}_{ij}^{r} - 1\big).
\end{aligned}
$$

□

Bibliography

Ale, B., L. Bellamy, R. Cooke, L. Goossens, A. Hale, A. Roelen, and E. Smith (2006). Towards a causal model for air transport safety - an ongoing research project. *Safety Science 44*(8), 657673.

Bedford, T. and R. Cooke (2002). Vines - A New Graphical Model for Dependent Random Variables. *Annals of Statistics 30*(4), 1031–1068.

Charniak, E. (1991). Bayesian networks without tears: making Bayesian networks more accessible to the probabilistically unsophisticated. *AI Magazine 12*, 50–63.

Cheng, J., D. A. Bell, and W. Liu (1997). An algorithm for Bayesian network construction from data. *Artificial Intelligence and Statistics*.

Chickering, D., D. Geiger, and D. Heckerman (1994). Learning Bayesian Networks is NP-Hard. *Technical Report MSR-TR-94-17, Microsoft Research*.

Conti, P. (1993). On some descriptive aspects of mesures of monotone dependence. *Metron 51*(3-4), 53–60.

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.*

Cooper, G. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning 9*, 309–347.

Cowell, R. (2005). Local Propagation in Conditional Gaussian Bayesian Networks. *The Journal of Machine Learning Research 6*, 1517 – 1550.

Cowell, R., A. Dawid, S. Lauritzen, and D. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Sciences. New York: Springer- Verlag.

Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis 93*, 40–57.

Edwards, W. (1998). Tools for and Experiences with Bayesian Normative Modeling. *Journal of the American Psychological Association 53*(4), 416–428.

Frank, M. (1979). On the simultaneous associativity of F(x,y) and x+y-F(x,y). *Aequationes Math. 19*, 194–226.

Friedman, N. and M. Goldszmidt (1996). Discretizing continuous attributes while learning Bayesian networks. In *In Proc. ICML*.

Ghosh, S. and S. Henderson (2002). Properties of the NOTRA method in higher dimensions. *Proc of the 2002 Winter Simulation Conference*, 263–269.

Goodman, L. A. and W. H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association 49*, 732–764.

Guo, H. and W. Hsu (2002). A survey of algorithms for real-time Bayesian network inference. In *AAAI/KDD/UAI Joint Workshop on Real-Time Decision Support and Diagnosis Systems, Canada*.

Hanea, A., 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., 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*.

Hanea, A., D. Kurowicka, R. Cooke, and D. Ababei (2007). Ordinal Data Mining with Non-Parametric Continuous Bayesian Belief Nets. *submitted to Computational Statistics and Data Analysis*.

Hanea, A. M. and D. Kurowicka (2008). Mixed Non-Parametric Continuous and Discrete Bayesian Belief Nets. *Advances in Mathematical Modeling for Reliability ISBN 978-1-58603-865-6*(IOS Press).

Heckerman, D. (1995). A tutorial on learning Bayesian networks. *Technical Report MSR-TR-95-06, Microsoft Research*.

Heckerman, D. and D. Geiger (1995). Learning Bayesian networks: a unification for discrete and Gaussian domains. *UAI*, 274284.

Heckerman, D., D. Geiger, and D. Chickering (1997). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning Journal 20(3)*.

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

Iman, R. and J. Helton (1985). A comparison of uncertainty and sensitivity analysis techniques for computer models. Technical report, NUREG/CR-3904 SAND84-1461 RG, Albuquerque.

Jaakkola, T., D. Geiger, and M. Jordan (1999). Variational probabilistic inference and the QMR-DT database. *Journal of Artificial Intelligence Research*.

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

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association 84*(405), 157–164.

Joe, H. (1990). Multivariate Concordance. *Journal of Multivariate Analysis 35*, 12–30.

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

John, G. and P. Langley (1995). Estimating continuous distributions in Bayesian classifiers. *UAI*, 338345.

Jordan, M., Z. Ghahramani, T. Jaakkola, and L. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning Journal 37*, 183233.

Kendal, M. G. (1945). The treatment of ties in ranking problems. *Biometrika 33*, 239–251.

Kendall, M. and J. Gibbons (1990). *Rank Correlation Methods*. First published in Great Britain 1948. Oxford University Press.

Kendall, M. and A. Stuart (1961). *The advenced theory of statistics*. London: Charles Griffin & Company Limited.

Kowalczyk, T. and M. Niewiadomska-Bugaj (2001). An algorithm for maximizing Kendal's τ. *Computational Statistics & Data Analisys 37*, 181–193.

Kurowicka, D. (2001). *Techniques in Representing High Dimensional Distributions*. PhD Dissertation, Delft Institute of Applied Mathematics.

Kurowicka, D. and R. Cooke (2004). Distribution - Free Continuous Bayesian Belief Nets. Proceedings Mathematical Methods in Reliability Conference.

Kurowicka, D. and R. Cooke (2006a). Completion problem with partial correlation vines. *Linear Algebra and its Applications*.

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

Kurowicka, D., J. Misiewicz, and R. Cooke (2000). Elliptical copulae. *Proc of the International Conference on Monte Carlo Simulation - Monte Carlo*, 209–214.

Lam, W. and F. Bacchus (1994). Learning Bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence 10*, 269–293.

Langseth, H. (2007). Bayesian Networks in Reliability Analysis. In *Invited Talk at the Mathematical Methods in Reliability Conference.*

Lauritzen, S. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association 87*, 10981108.

Lauritzen, S. (1996). *Graphical Models.* Clarendon Press, Oxford.

Lauritzen, S. and F. Jensen (2001). Stable local computation with conditional Gaussian distributions. *Statistics and Computing 11*, 191203.

Lawley, D. and M. Maxwell (1963). *Factor Analysis as a Statistical Method.* London: Butterworths Mathematical Texts.

MacKay, D. (1999). An introduction to Monte Carlo methods. *Learning in Graphical Models Cambridge, MA,.*

Margaritis, D. (2005). Distribution-Free Learning of Bayesian Network Structure in Continuous Domains. Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh.

Mesfioui, M. and A. Tajar (2005). On the properties of some nonparametric concordance measures in the discrete case. *Journal of Nonparametric Statistics 17*, 541–554.

Micheas, A. and K. Zografos (2006). Measuring stochastic dependence using φ-divergence. *Journal of Multivariate Analysis 97*, 765784.

Morales, O., D. Kurowicka, and A. Roelen (2007). Eliciting conditional and unconditional rank correlations from conditional probabilities. *Reliability Engineering and System Safety*. doi: 10.1016/j.ress.2007.03.020.

Morales-Napoles, O., D. Kurowicka, R. Cooke, and D. Ababei (2007). Continuous-discrete distribution free Bayesian belief nets in aviation safety with UniNet. *Technical Report TU Delft*.

Morgenstern, R., W. Harrington, J. Shis, R. Cooke, A. Krupnick, and M. Bell (2008). Accountabilty Analysis of Title IV of the 1990 Clean Air Act Amendments. An Approach using Bayesian Belief Nets. In *Poster.*

Murphy, K. (2002). An introduction to graphical models. Technical report, Intel Research.

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

Neslehova, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis 98*(3), 544–567.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufman Publishers.

Pearl, J. and T. Verma (1991). A theory of inferred causation. *KR'91: Principles of Knowledge Representation and Reasoning*, 441–452.

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

Ramsey, P. (1989). Critical values for Spearman's rank order correlation. *Journal of educational statistics 14*(3), 245–253.

Renyi, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hungar. 10*, 441451.

Roelen, A., R. Wever, A. Hale, L. Goossenes, R. Cooke, R. Lopuhaa, M. Simons, and P. Valk (2004). Casual modeling using Bayesian belief nets for integrated safety at airports. *Risk Decision and Policy 9*(3), 207–222.

Schmid, F. and R. Schmidt (2007). Multivariate extensions of Spearman's rho and related statistics. *Statistics & Probability Letters 77*, 407416.

Shachter, R. and C. Kenley (1989). Gaussian influence diagrams. *Management Science 35*(5), 527–550.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology 15*, 72–101.

Spearman, C. (1906). A footrule for measuring correlation. *British Journal of Psychology 2*, 89–108.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.

"Student" (1921). An experimental determination of the probable error of dr. Spearman's correlation coefficients. *Biometrika 13*, 263–282.

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

Vandenhende, F. and P. Lambert (2000). Modeling repeated ordered categorical data using copulas. In *Institute of Statistics, Universite Catholique de Louvain, Belgium, Discussion paper*.

Vandenhende, F. and P. Lambert (2003). Improved rank-based dependence measures for categorical data. *Statistics and Probability Letters 63*, 157–163.

Whittaker, J. (1990). *Graphical Models in applied multivariate statistics*. Chichester: John Wiley and Sons.

Wright, S. (1921). Correlation and causation. *Jour. Agric. Res. 20*, 557–585.

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

Zhang, N. and D. Poole (1994). A simple approach to Bayesian network computations. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence*.

Summary

High dimensional probabilistic modelling using graph theory is employed in several scientific fields. Graphical models proved to be a flexible probabilistic framework, and their use has increased substantially. They merge graph theory and probability theory to provide a general setting for models in which a number of variables interact. There are two main types of graphical models: directed and undirected.

Our focus is on the directed graphical models called Bayesian Belief Nets (BBNs). A BBN encodes the probability density or mass function of a set of variables by specifying a set of conditional independence statements in a form of an acyclic directed graph and a set of probability functions. It provides a simple way to visualize the structure of a probabilistic model. Until recently BBN models were restricted to structures containing discrete and/or Gaussian variables. Uncertainty distributions may not be assumed to conform to any parametric form. Algorithms for specifying, sampling and analysing high dimensional distributions should therefore be non-parametric. This thesis proposes a number of algorithms for non-parametric BBNs.

Chapter 1 contains a short overview of the *classical* BBN models and discusses their disadvantages.

Chapter 2 reviews the details of non-parametric BBNs using the copula-vine modelling approach and introduces two new methods. The first one is a hybrid approach, that combines the reduced assessment burden of the continuous BBNs with the fast updating algorithms of their discrete counterparts. The drawbacks of this method are discussed, and these provide the motivation of introducing a second method. A new sampling protocol based on the normal copula is proposed. Normal vines are used to realize the dependence structure specified via (conditional) rank correlations on the continuous BBN.

The latter approach is extended to include ordinal discrete random variables. In contrast with the continuous case, the rank correlation of two discrete variables and the rank correlation of their underlying uniforms are not equal. Therefore we first study the relationship between these two rank correlations. Chapter 3 presents a generalisation of the population version of Spearman's rank correlation for the case of ordinal discrete random variables.

In Chapter 4 we present two large ongoing projects in which mixed non-parametric continuous & discrete BBNs are the tool used in the analysis.

Chapter 5 is concerned with non-parametric BBNs from a different perspective, namely as a tool for mining ordinal multivariate data. We propose a method for learning a BBN from data. The main advantage of this method is that it can handle a large number of continuous variables, without making any assumption about their marginal distributions. The learning procedure is fast and flexible. Once we have learned the BBN from data, we can further use it for prediction or diagnosis by employing the methods described in the previous chapters. We illustrate the method proposed using a database of pollutants emissions and fine particulate concentrations.

In Chapter 6 the most important results of this work are summarised and conclusions are formulated.

Most of the methods discussed in this thesis are implemented in the software application UniNet. A short description of UniNet is given in Chapter 7.

Samenvatting

Multi-dimensionale waarschijnlijkheidsmodellen die gebaseerd zijn op grafentheorie worden in verschillende wetenschappelijke terreinen gebruikt. Deze op grafen gebaseerde modellen hebben bewezen een flexibel denkkader te zijn binnen de kansrekening, en hun gebruik is fors toegenomen. Ze verenigen grafentheorie met kansrekening, en bieden daarmee een algemene basis voor modellen waarin verschillende variabelen interacteren. Er zijn twee hoofdtypen van grafische modellen, gerichte en ongerichte.

Wij concentreren ons op de gerichte grafische modellen die bekend staan als Bayesiaanse Belief Netwerken (BBNs). Een BBN codeert de kansdichtheid, of de kansmassafunctie, van een verzameling variabelen door een verzameling voorwaardelijke onafhankelijkheidsrelaties als een acyclische gerichte grafe weer te geven, met daarbij een verzameling kansverdelingen. Het is een eenvoudige manier om de structuur van een stochastisch model weer te geven. Tot voor kort waren BBN modellen beperkt tot structuren met alleen discrete en/of Gaussische variabelen. Onzekerheidsverdelingen mogen echter niet worden aangenomen zich aan enige parametrische vorm te conformeren. Algoritmen die bedoeld zijn om multidimensionale verdelingen te analyseren en beschrijven zouden dus parametervrij moeten zijn. Dit proefschrift stelt een aantal algoritmen voor voor parametervrije BBNs.

Hoofdstuk 1 bevat een overzicht van de *klassieke* BBN modellen en bespreekt hun nadelen.

Hoofdstuk 2 vat de details van parametervrije BBNs samen met behulp van de copula-vine benadering en introduceert twee nieuwe methodes. De eerste is een hybride benadering die de kleinere assessment last van continue BBNs combineert met de snelle updating algoritmen van de discrete BBNs. De nadelen van deze aanpak worden besproken, en ze vormen een motivering voor het introduceren van een tweede methode. Een nieuw simulatieprotocol wordt voorgesteld, gebaseerd op de normale copula. Hierbij worden normale vines gebruikt om de afhankelijkheidsstructuur, die wordt gespecificeerd door (voorwaardelijke) rangcorrelaties op de continue BBN, te realiseren.

De laatstgenoemde aanpak wordt uitgebreid om ook ruimte te bieden aan ordinale discrete stochastische variabelen. In tegenstelling tot het continue geval zijn de rangcorrelaties van twee discrete variabelen en die van de onderliggende

uniforme stochasten niet gelijk, daarom bestuderen we eerst de relatie tussen deze twee rangcorrelaties. Hoofdstuk 3 stelt een generalisatie voor van de populatieversie van Spearman's rangcorrelatie voor het geval van ordinale discrete stochastische variabelen.

In Hoofdstuk 4 presenteren wij twee grote lopende projecten waarin gemengde parametervrije continue en discrete BBNs worden gebruikt voor de analyse.

Hoofdstuk 5 gaat hoofdzakelijk over parametervrije BBNs vanuit een ander perspectief, namelijk als gereedschap voor het analyseren van ordinale multivariate data. Het grootste voordeel van deze methode is dat hij grote aantallen continue variabelen kan verwerken, zonder aannamen te hoeven doen over hun marginale verdelingen. De voorgesteld algoritme voor het leren van Bayesiaanse netwerken uit data is snel en eenvoudig. Als we eenmaal de BBN van de data hebben gevonden, dan kunnen we deze gebruiken voor predictie en diagnostiek door de methoden uit de voorgaande hoofdstukken toe te passen. We illustreren deze methode aan de hand van een database over vervuilingsuitstoot en fijnstofconcentraties.

In Hoofdstuk 6 worden de belangrijkste bevindingen van dit werk samengevat en geformuleerd. De meeste van de in dit proefschrift besproken methoden worden toegepast in het computerprogramma UniNet. Een korte beschrijving van UniNet word gegeven in Hoofdstuk 7.

Acknowledgements

I am very lucky to be able to say, and proud to admit that I did not go through the experience of being a *young researcher* at the Delft University of Technology all by myself. There are many people that have influenced, helped and supported me all these years, who must be thanked.

I shall follow an unwritten rule and begin with my thoughts for the people that were there for me from a professional point of view. I would like to express my gratitude to my supervisor Roger Cooke, first of all for offering me the possibility to be a PhD student under his supervision. His ideas combined with his unique kind of enthusiasm gave contour to this thesis. Without his unwavering trust that our methods will "take over the world" my time as a PhD student would have been a lot duller. I must also acknowledge Dorota Kurowicka for her ideas, suggestions and her sharp critical point of view about some of my writings.

A very special thanks goes out to Jolanta Misiewicz for her time spent with me in Zielona Gora, and for her infinite patience in checking my horrible calculations. I would also like to thank the members of my committee for their comments and ideas shared and discussed during workshops, conferences, or even lunch breaks.

Appreciation also goes out to Carl and Cindy for all the instances in which their assistance helped me along the way. A series of far too many names to mention separately is that of the master students from our group and roommates that I have worked, talked, lunched and had fun with over the years. My gratitude goes out to all of them.

Shifting my attention onto more personal acknowledgements, I would like to thank the people that are dearest to me: my friends[6] and my family.

Writing this thesis, being where I am now, as happy as I am now, would not have been possible if it had not been for my brother and sister[7]. Thank you Remus and Dana for being marvellous siblings and friends, for giving me the most honest, proper, helpful, and trustworthy advices and solutions to all my personal and professional problems. Remus, I hope everybody can be blessed with a brother like you.

[6]I know I do not need to list your names, you know exactly who you are!

[7]In law.

That brings me to my mum and dad, and I would write the following in Romanian if somebody else did not have this idea before me. I can not omit my dad from the list of people that had a big influence upon my life. He mostly had a good influence and I want to thank him for that. The bad influence that translated in little professional self-confidence was "fixed" by other people; the most important of these being Roger Cooke. Thank you Roger for your trust[8].

Words fail if I try to express my enormous gratitude for my mum. Her constant support and encouragement on all fronts, her friendship and her faith in me kept me going with a smile on my face, a smile that tries to imitate hers[9].

This is the moment when I would really like to be a writer, rather than a mathematician. Because I do not master the words, I have to express my feelings for the most important person in my life, in a simple way. Dan has been always there for me, as I turned from a shy first year student with two braids into the author of this PhD thesis. Thank you Dan for making everything beautiful, special and possible. Thank you for being the most important part of my story.

A wholehearted *thank you* to all.

Delft, *Anca Hanea*
December 2008

[8] You get to be mentioned twice.

[9] Când încerc să'mi exprim recunoştinţa pentru mama, cuvintele nu mai fac faţa. Sprijinul şi încurajările ei constante, pe toate fronturile, prietenia şi încrederea ei în mine m'au făcut să merg înainte cu zâmbetul pe buze, un zâmbet care încearcă să'l imite pe al ei.

Curriculum Vitae

Anca Maria Hanea was born in Bucureşti, România on March 5, 1979. From 1993 she attended secondary school at *Colegiul National Sf. Sava* in Bucureşti, where she graduated in 1997.

In the same year she started her studies at the Department of Mathematics, the Faculty of Mathematics, University of Bucharest and received her Bachelor's degree in 2001. After graduation she joined as a staff member of the Department of Mathematics at the Technical University of Civil Engineering, Bucharest. She continued the research work at the Technical University of Civil Engineering, until August 2002.

In September 2002 Anca moved to the Netherlands, where she followed a Master of Science Programm in Risk and Environmental Modelling at Delft University of Technology. In June 2004, Anca defended her Master thesis: "Decision Support Systems for Oil Spills in The Wadden Sea" and obtained the M.Sc. degree (magna cum laude) in Applied Mathematics.

From September 2004 Anca Hanea followed a Ph.D. research program at the Department of Applied Mathematics, Delft University of Technology, under the supervision of Prof. Dr. R.M. Cooke. This research was financially supported by the Netherlands Organisation for Scientific Research (NWO) in the framework of the "Copulae for High Dimensional Distributions" project 613.002.052. The work which has been done from September 2004 until September 2008 results in the thesis entitled "Algorithms for Non-Parametric Bayesian Belief Nets", to be defended on the 15th December 2008 in Delft.

Anca Hanea is currently employed as a researcher by the Delft University of Technology.