Chapter 1

Non-Parametric Bayesian Belief Nets versus Vines

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

This chapter reviews aspects of non-parametric Bayesian belief nets (NPBBN). The theory behind NPBBNs is closely related to that of regular vines and it benefits from the latter's developments. It also offers an alternative to undirected graphical models in general, and to regular vines in particular. The differences and similarities in modelling using directed versus undirected graphs are discussed in this chapter from the perspective of NPBBNs and vines. Until recently, Bayesian belief nets (BBNs) were either discrete or discrete-normal. Despite their popularity, both suffer from severe limitations. Discrete BBNs are limited by size and complexity, discrete-normal BBNs are limited by the assumption of joint normality. NPBBNs were introduced to overcome these limitations. Algorithms for specifying, sampling and analysing high dimensional distributions using NPBBNs are developed and successfully applied in decision support systems.

Contents

1.1. Introduction or: how to represent information burdened by uncertainty

Understanding and representing multivariate distributions along with their dependence structure is a highly active area of research. A large body of scientific work treating multivariate models is available. This chapter in particular, and this book in general, advocates graphical models to represent high dimensional distributions with complex dependence structures.

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.

There are two main types of graphical models: directed, based on directed acyclic graphs (DAGs), and undirected, generally referred to as Markov networks. The regular vines are a generalisation of Markov trees, hence they fall into the former category, whereas the Bayesian belief nets (BBNs) belong to the latter. Why or when to use one graphical model or another is not a question with a straightforward answer. This chapter will provide some insights into the differences and similarities between the two types of models, and hopefully these will serve as guidelines for modelers.

Both directed and undirected models consist of a qualitative and a quantitative part. The qualitative part is represented by the graph itself together with the (in)dependence relationships entailed by it. Maybe the most important difference between directed and undirected graphs, in general, is that they make different statements of conditional independence. We will first focus on the differences arising from the graphical structures, rather than the quantification of a joint multivariate distribution.

The absence of a link between two nodes means that any dependence between these two variables is mediated via some other variables, hence they encode conditional (in)dependence statements between variables. Given the nested tree structure of a regular vine, one can consider them as fully connected graphs. In this sense, in regular vines, the concept of conditional independence is weakened to allow for various forms of conditional dependence. Is this an advantage or a disadvantage of vines? The answer depends on many factors, which may lead to the conclusion that the question is ill-posed.

A number of examples will shed some light on the matter. Consider 3 random variables X_1, X_2 and X_3 represented as nodes in a graphical structure. The node that corresponds to variable X_i is denoted by i. Let node 3 have converging links. This is a configuration that yields conditional independence in Markov networks and conditional dependence in BBNs. The structure in Fig. 1.1(a) entails the *conditional dependence* of X_1 and X_2 given X_3[a], whereas Fig. 1.1(b) entails the *conditional independence* of X_1 and X_2 given X_3.

[a] and the independence of X_1 and X_2.

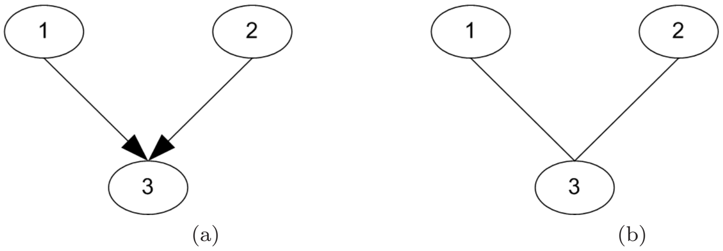

Fig. 1.1. Node with converging links. (a) Node with converging arrows in a BBN. (b) Node with converging edges in a Markov network.

The possibility of representing the combination of statements in Fig. 1.1(a) may be regarded as an advantage of BBNs over undirected structures, since it permits the display of induced and non-transitive dependencies. This configuration also represents the main difference between the separation properties in the directed and undirected graphs. In directed graphs, the direction-dependent criterion of connectivity called the *d-separation criterion* consists in the above rule for converging arrows, plus the usual cutset criterion of Markov networks, whenever the arrows are diverging or cascaded [16]. If two nodes of a BBN are d-separated by a set of nodes, then the corresponding variables are conditionally independent, given that set.

Remark 1.1. If two nodes are not d-separated it does not necessarily mean that the corresponding variables are not conditionally independent. In other words, whenever an arc or an unblocked path[b] exists between two nodes, it is not necessarily the case that the corresponding variables are dependent.

Regular vines however may be also used to represent the independence of X_1 and X_2, and the conditional dependence of X_1 and X_2 given X_3. Nevertheless, the graphical structure alone will not suffice in completing this task and this might be viewed as a disadvantage of regular vines. Given the full connectivity of vines (conditional) dependencies and/or independencies can only be represented through quantification. Edges of a regular vine can be associated with (conditional) rank correlations. If these rank correlations are realised by copulae with the zero independence property, representing the independence of X_1 and X_2 reduces to associating the edge between

[b] Intuitively, an unblocked path *may* carry information, or dependence between end nodes. For exact definitions we refer to [16].

them with a zero rank correlation. This is shown in Fig. 1.2(a). Yet, the conditional dependence of X_1 and X_2 given X_3 is not obvious. A few calculations are needed in order to verify that, and a different graph is needed to actually visualise it. Fig. 1.2(b) shows a non-zero conditional rank correlation between X_1 and X_2 given X_3, but fails to represent the independence of X_1 and X_2.

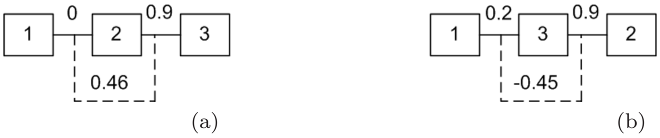

Fig. 1.2. D-vines "representing" induced and non-transitive dependencies. (a) D-vine representing the independence of X_1 and X_2. (b) D-vine representing the conditional dependence of X_1 and X_2 given X_3.

It is worth remembering that the present discussion regards solely the representation of certain (conditional) (in)dependencies using different graphical structures, and not the full representation/quantification of joint distributions. The specification of (conditional) rank correlations on the edges of a regular vine serves here this purpose only.

Another feature of BBNs that can be regarded as an advantage over regular vines is that conditional independencies are represented by missing arcs, therefore by deleting arcs certain conditional independencies become visible in the graph. Consider the D-vine on 4 variables in Fig. 1.3. In this example and further in this chapter, the copulae used to realise the (conditional) rank correlations associated to the edges of a regular vine will posses the zero independence property.

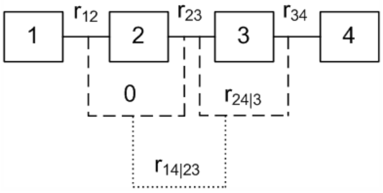

Fig. 1.3. A D-vine on 4 variables representing the following: $X_1 \perp X_3|X_2$; $X_2 \not\perp X_4|X_3$; $X_1 \not\perp X_4|\,(X_2, X_3)$; $X_1 \not\perp X_2$, $X_2 \not\perp X_3$, $X_3 \not\perp X_4$.

Variables X_1 and X_3 are independent given X_2. Independence is denoted by $\perp$, e.g. $X_1 \perp X_3|X_2$. The notation $X_2 \not\perp X_4|X_3$ means that X_2

and X_4 are not conditionally independent given X_3. If a (conditional) rank correlation from the D-vine is not replaced by zero the corresponding variables are considered to be (conditionally) dependent. The information represented by the D-vine in Fig. 1.3 can be represented using a saturated BBN, from which the arc between X_1 and X_3 is deleted. In this way the dependence between the two variables is mediated only via X_2 (see Fig. 1.4(a)). Further, X_2 and X_4 are conditionally dependent given X_3. Since the presence of arcs does not guarantee dependence between variables (see Remark 1.1), this statement cannot be represented with a BBN. The best one could do is to avoid representing the opposite (i.e. $X_2 \perp X_4|X_3$). The dependence between X_2 and X_4 is not mediated only through X_3, therefore the arc between them can be deleted (see Fig. 1.4(b)). This of course will introduce a new conditional independence statement, i.e. $X_2 \perp X_4|\,(X_1, X_3)$, but it will not necessarily violate the requirements imposed by the D-vine. The resultant structure is presented in Fig. 1.4(c).

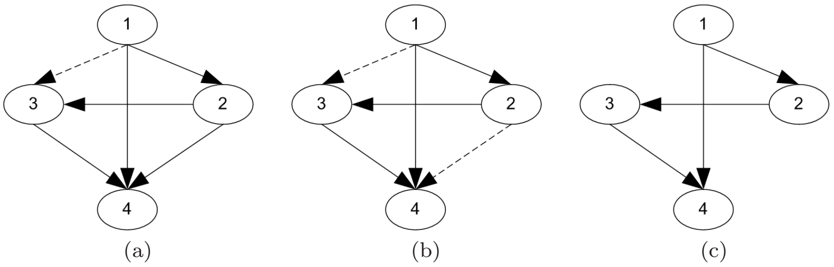

Fig. 1.4. (a) A BBN with 4 nodes and 5 arcs representing $X_1 \perp X_3|X_2$. (b) A BBN with 4 nodes and 4 arcs representing $X_1 \perp X_3|X_2$. (c) The same BBN as in b).

Only 4 arcs are necessary in order to represent the same conditional independence statements as in the D-vine. The reduction in the number of arcs constitutes a major advantage, since it results in a sparser, more readable structure. Another example of a set of conditional independence statements represented with a D-vine with 15 edges versus a BBN with 6 arcs is presented in Fig. 1.5 and Fig. 1.6.

Following the same strategy as before, i.e. starting with the saturated BBN and removing the arcs corresponding to the independence statements, results in the BBN in Fig. 1.6(a). As expected, the number of arcs is reduced to 11. Nevertheless, if one only wants to preserve the conditional independence statements shown in the regular vine and not to violate the

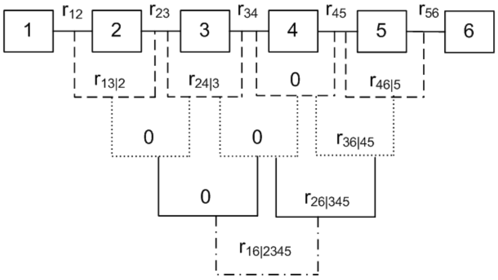

Fig. 1.5. A D-vine on 6 variables representing 4 conditional independence statements.

conditional dependencies, the structure can be reduced even further, e.g. Fig. 1.6(b).

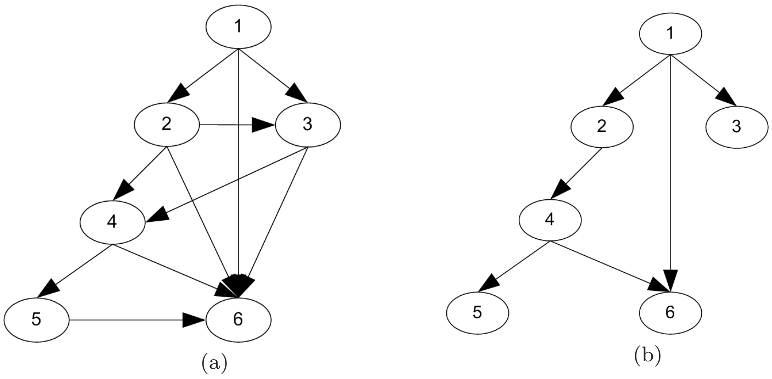

Fig. 1.6. (a) BBN with 6 nodes and 11 arcs representing the same conditional independence statements as the D-vine in Fig 1.5. (b) BBN with 6 nodes and 6 arcs representing the same conditional independence statements as the D-vine in Fig 1.5.

In larger structures, with many conditional independence statements present, the reduction might be even more dramatic. Nevertheless, there are configurations in which deleting arcs from a saturated BBN (corresponding to a regular vine) does not result in a better "picture". Consider the D-vine in Fig. 1.7.

Starting with the saturated BBN and deleting the arcs between X_1,X_3 and X_2, X_4 will result in the BBN in Fig. 1.4(c). But, in this structure X_2 and X_4 are not d-separated by X_3. This does not imply that they are not conditional independent given X_3. They might be, but this conditional independence is not visible anymore, and the BBNs' advantage of being

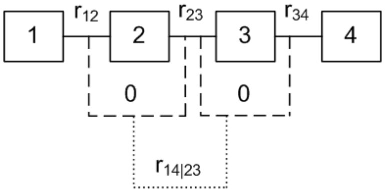

Fig. 1.7. A D-vine on 4 variables representing the following: $X_1 \perp X_3|X_2$; $X_2 \perp X_4|X_3$; $X_1 \not\perp X_4|\,(X_2, X_3)$; $X_1 \not\perp X_2$, $X_2 \not\perp X_3$, $X_3 \not\perp X_4$.

visually more intuitive vanishes. Any reorientation of the arcs will fail to represent - via d-separation - both conditional independence statements.

On the other hand, starting with the BBN structure in Fig. 1.4(c) (rearranged as in Fig. 1.8(a)) and trying to represent its conditional independencies with a vine might prove difficult. Fig. 1.8(a) encodes $X_4 \perp X_2|X_1, X_3$ and $X_1 \perp X_3|X_2$. To represent the first statement on a D-vine, variable X_1 has to be before variable X_2 in the first tree (see Fig. 1.8(b)), whereas to represent the second statement the order of these variables has to change (see Fig. 1.8(c)).

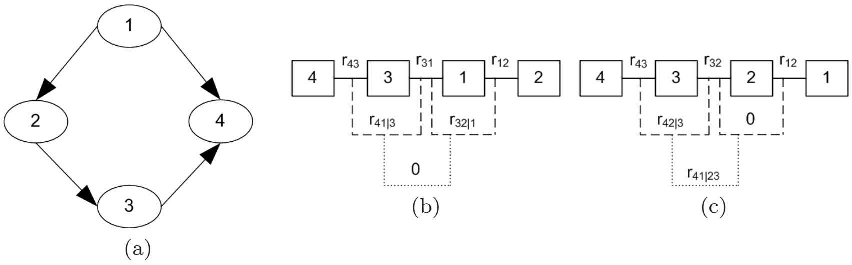

Fig. 1.8. (a) A BBN representing $X_4 \perp X_2|X_1, X_3$ and $X_1 \perp X_3|X_2$. (b) A D-vine representing $X_4 \perp X_2|X_1, X_3$. (c) A D-vine representing $X_1 \perp X_3|X_2$.

The choice between representing a multivariate distribution using a regular vine, or using a BBN depends on many factors. A few of them, related exclusively to the graphical representation of (in)dependence statements were discussed above. Other factors will be explored throughout this chapter.

The rest of the chapter is organized as follows. We first introduce nonparametric Bayesian belief nets (NPBBNs) and their connection with regular vines. Differences in sampling and performing inference using a NPBBN versus using a regular vine are further discussed. The issues of model learning and validation are addressed and some applications of the NPBBNs methodology are finally presented. The last section gathers conclusions.

1.2. Non-Parametric Bayesian Belief Nets: sampling and conditionalising

This chapter concentrates on BBNs. As already mentioned, BBNs are DAGs, whose nodes represent univariate random variables and arcs represent direct influences[c].

The origin of BBNs can be tracked back in the early decades of the 20th century to the pioneering work of Sewell Wright [19] who developed path analysis to help the study of genetic inheritance.

In their most popular form, BBNs were introduced in the 80's as a knowledge representation formalism to encode and use the information acquired from human experts in automated reasoning systems to perform diagnostic and prediction [16].

BBNs provide a compact representation of high dimensional distributions of a set of variables and encode their joint density/mass function by specifying a set of conditional independence statements and a set of probability functions. The graph itself and the (conditional) independence relations that are entailed by it form the qualitative part of a BBN model. The quantitative part of the model consists of the conditional probability functions associated with the variables. In Section 1.1 we concentrated our attention on the qualitative part of BBNs. Further we will mainly discuss their quantitative aspects and the techniques to build high dimensional distributions.

Until recently, BBNs were discrete, normal or discrete-normal. In discrete BBNs nodes represent discrete random variables. These models specify marginal distributions for source nodes, and conditional probability tables for child nodes. If the nodes of a BBN correspond to variables that follow a joint normal distribution, we talk of Gaussian BBNs (or normal BBNs) [16, 18]. 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 al-

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

gebraically independent manner [18].

Despite their popularity, they suffer from severe limitations. Discrete BBNs are limited by size and complexity; normal and discrete-normal BBNs are limited by the assumption of joint normality[d].

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. Regular vines allow us to move beyond discrete BBNs without defaulting to the joint normal distribution. When no marginal distribution assumption is made, we talk of non-parametric BBNs, abbreviated NPBBNs. NPBBNs and their relationship with regular vines were introduced in [11] and extended in [5]. The focus of this section is on quantifying and building a joint distribution using a NPBBN.

A *NPBBN* is a DAG, together with a set of (conditional) rank correlations, a copula class parametrised by the rank correlation, with the zero independence property, and a set of marginal distributions. In NPBBNs nodes are associated with arbitrary distributions and arcs with (conditional) rank correlations that are realised by the chosen copula. In continuous NPBBNs nodes are associated with continuous invertible distribution functions. The nodes of a NPBBN will be assumed continuous unless otherwise specified. Further in this chapter, whenever we speak of NPBBNs, we mean the DAG together with the specification of rank correlations, copula and margins.

The DAG of a NPBBN induces a (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 $Pa(i)$.

Each variable is associated with a conditional probability function of that variable given its parents in the graph, $f_{i|Pa(i)}, i = 1, \ldots, n$. The conditional independence statements encoded in the graph allow us to write[e]:

$$f_{1,2,\ldots,n} = \prod_{i=1}^{n} f_{i|Pa(i)}. \tag{1.1}$$

[d] For a detailed discussion about the disadvantages of discrete and normal BBNs we refer to Chapter 1 of [4].

[e] This factorisation is of course valid for BBNs in general and not only for NPBBNs.

For each variable i with parents $i_1...i_{p(i)}$, we associate the arc $i_{p(i)-k} \longrightarrow i$ with the conditional rank correlation:

$$\begin{cases} r_{i,i_{p(i)}}, & k=0 \\ r_{i,i_{p(i)-k}|i_{p(i)},...,i_{p(i)-k+1}}, & 1 \leq k \leq p(i)-1. \end{cases} \tag{1.2}$$

The assignment is vacuous if $\{i_1...i_{p(i)}\} = \emptyset$ (see Fig. 1.9).

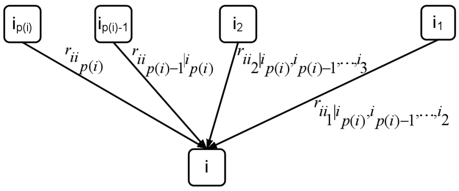
$i_{p(i)}$
$i_{p(i)-1}$
i_2
i_1
$r_{ii_{p(i)}}$
$r_{ii_{p(i)-1}|i_{p(i)}}$
$r_{ii_2|i_{p(i)},i_{p(i)-1},...,i_3}$
$r_{ii_1|i_{p(i)},i_{p(i)-1},...,i_2}$
i

Fig. 1.9. Node i of a NPBBN and the set of parent nodes for i.

Therefore, every arc in the NPBBN is assigned a (conditional) rank correlation between parent and child. These assignments are made according to a protocol presented in [11]. The conditional rank correlations need not be constant, although they are taken to be constant in the following example[f]. We will illustrate the protocol for assigning (conditional) rank correlations to the arcs of a NPBBN with an example.

Example 1.1. Let us consider the undirected cycle on 4 variables in Fig. 1.10. This structure is similar with the structure presented in Fig. 1.8(a).

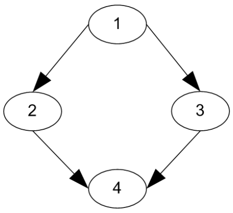

Fig. 1.10. BBN with 4 nodes and 4 arcs.

The DAG of this NPBBN induces 2 orderings[g] of the variables: 1, 2, 3, 4, or 1, 3, 2, 4. Let us choose 1, 2, 3, 4. The factorization of the joint distribution is:

[f] The conditional rank correlations must be constant when the normal copula is used.
[g] Such an ordering of the variables is referred to as *sampling order* or *topological order*.

$$P(1)P(2|1)P(3|1\underline{2})P(4|23\underline{1}). \tag{1.3}$$

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 NPBBN. The correlation between the child and its first parent[h] will be an unconditional rank correlation, and the correlations between the 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 NPBBN in Fig. 1.10 is: $\{r_{21}, r_{31}, r_{42}, r_{43|2}\}$. For each term i ($i = 1, \ldots, 4$) of the factorization (1.3) 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. Fig. 1.11 shows the D-vines built for variables 2, 3, 4.

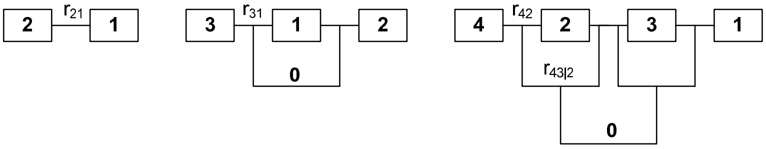

Fig. 1.11. $\mathcal{D}^2, \mathcal{D}^3, \mathcal{D}^4$ for Example 1.1.

Building the D-vines is not a necessary step in specifying the rank correlations[i], but it is essential in proving a result that not only establishes the connection between NPBBNs and vines, but is also crucial for the development of NPBBNs. The result will be further formulated; for its proof we refer to [5]:

Given a continuous NPBBN on n variables, the joint distribution of the variables is uniquely determined. This joint distribution satisfies the characteristic factorization (1.1) and the conditional rank correlations in (1.2) are algebraically independent.

The (conditional) rank correlations and the marginal distributions needed in order to specify the joint distributions represented by the NPBBN can be retrieved from data if available, or elicited from experts [14].

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

[i] These are assigned directly to the arcs of the BBN. Each arc is associated with a (conditional) parent-child rank correlation as in Fig. 1.9.

1.2.1. *Sampling a NPBBN*

Since no analytical/parametric form of the joint distributions is available, the only way to stipulate it is by sampling it. In order to sample a NPBBN we will use the procedures for regular vines presented in Chapter 3. Variable X_i is sampled using the procedure for the vine $\mathcal{D}^i$. When using regular vines to sample a continuous NPBBN, it is not in general possible to keep the same order of variables in successive 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. If the order of variables does not change from one vine to another the sampling procedure for the NPBBN coincides with the sampling procedure for the regular vine built for the last variable in the ordering (for details and examples see [13]). In Fig. 1.11, 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 NPBBN cannot be represented as just one D-vine. This particularity of an undirected cycle was already noticed in Fig. 1.8 from Section 1.1. In order to sample X_4 we use the sampling procedure described in Chapter 3 of this book:

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

which, using the conditional independencies from the graph, reduces to:

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

The conditional distribution $F_{r_{32};x_2}(x_3)$ is not given explicitly, but it can be calculated as follows:

$$F_{3|2}(x_3) = \int_0^{x_3} \int_0^1 c_{21}(x_2, x_1) c_{31}(v, x_1) dx_1 dv,$$

where c_{i1} is the density of the chosen copula with correlation r_{i1}, $i \in \{2, 3\}$.

For each sample, one needs to calculate the numerical value of the double integral. In this particular case, when only one double integral needs to be evaluated, it can be easily done without excessive computational burden. If the NPBBN contains an undirected cycle of five variables, and the same sampling procedure is applied, a triple integral will have to be calculated. The bigger the undirected cycle is, the larger the number of multiple integrals that have to be numerically evaluated.

In large structures, that contain large undirected cycles, this may constitute a big disadvantage of NPBBN in comparison with vines. If the multivariate distribution can be represented and assessed using one single regular vine, no extra calculations are needed in order to obtain samples from the joint distribution, hence the computational time reduces drastically.

Nevertheless, the disadvantage mentioned above vanishes when the normal copula is used. A different sampling protocol based on the normal copula uses the properties of normal vines to realise the dependence structure specified via (conditional) rank correlations on the NPBBN. This sampling protocol is presented in Chapter 3. The main advantage of this method is that everything is calculated on the joint normal vine, hence we can reorder the variables (if necessary) and recompute all partial correlations needed. This results in a dramatic decrease in the computational time. For examples and comparisons see [5].

It is worth mentioning that the approach to continuous NPBBNs using vines is extended to include ordinal discrete random variables. The dependence structure in the NPBBN is defined via (conditional) rank correlations, hence with respect to the underlying uniform variables. The rank correlation of 2 discrete variables and the rank correlation of their underlying uniforms are not equal. The relationship between them is established in [6]. This relationship is based on a generalisation of the population version of Spearman's rank correlation coefficient for the case of ordinal discrete random variables.

Since the sampling procedure for NPBBNs is based on the one for regular vines, we cannot talk about the advantages of the former compared to the latter.

1.2.2. *Conditionalising a NPBBN*

Maybe one of the most important features of probabilistic graphical models is that they can be used for inference. One can calculate the distributions of unobserved nodes, given the values of the observed ones, i.e. conditional distributions.

For regular vines, if values of some variables are observed, the results of sampling the model - conditional on these values - can be obtained either by sampling again the structure (the cumulative approach), or by using the density approach, both presented in Chapter 3. The new conditional distribution, although calculated, cannot be easily visualised and compared with the unconditional one. Even if this is merely an implementation issue for graphical software, still NPBBNs hold the advantage that conditionalisation can be visualised and interpreted in terms of the directionality of arcs. In other words, if the reasoning is done "bottom-up" (in terms of the directionality) the NPBBN is used for diagnosis, whereas if it is done "top-down", the NPBBN serves for prediction. Following the principle *a picture is worth a thousand words*, we will continue with an example. This example is loosely based on an ongoing project undertaken by the European Union that uses the NPBBNs' methodology. The name of the project is Beneris (which stands for Benefit and Risk) and it focuses on the analysis of health benefits and risks associated with food consumption[j]. The model introduced here is a highly simplified version of the NPBBN model used in the project [10]. The goal is to estimate the beneficial and harmful health effects in a specified population, as a result of exposure to various contaminants and nutrients through ingestion of fish.

Example 1.2. Fig. 1.12(a) resembles the version of the model that we are considering for purely illustrative purposes.

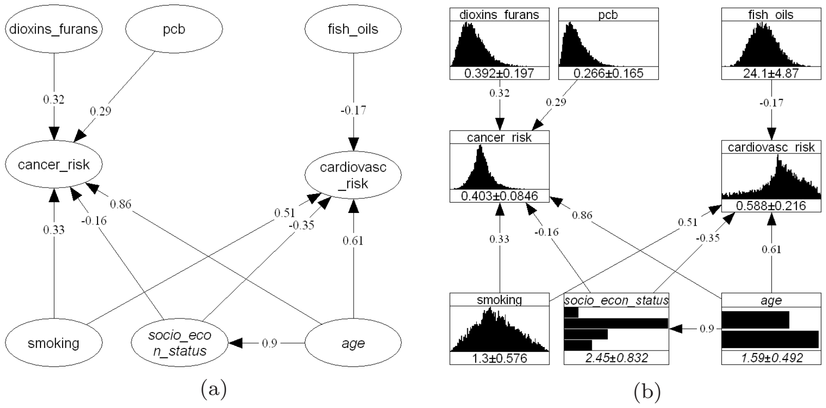

Fig. 1.12. (a) Simplified fish consumption NPBBN. (b) Simplified fish consumption NPBBN with histograms.

The variables of interest for this model are the health endpoints resulting from exposure to fish constituents, namely cancer and cardiovascular risk. These risks are defined in terms of remaining lifetime risks. The 3 fish constituents that are considered are: dioxins/furans, polychlorinated biphenyls, and fish oil. The first two are persistent and bio-accumulative toxins which cause cancer in humans. Fish oil is derived from the tissues of oily fish and has high levels of omega-3 fatty acids which regulate

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

cholesterol and reduce inflammation throughout the human body. Personal factors such as smoking, socioeconomic status and age may also influence cancer and cardiovascular risk. Smoking is measured as yearly intake of nicotine during smoking and passive smoking, while the socioeconomic status is measured by income, and is represented by a discrete variable with 4 income classes (unemployed, blue collars, white collars, and farmers and entrepreneurs) . The age is taken, in this simplified model, as a discrete variable with 2 states, 15 to 34 years, and 35 to 59 (we are considering only a segment of the whole population).

The distributions of the variables are presented in Fig. 1.12(b) together with their means and standard deviations. They are chosen by the author for illustrative purposes only. So are the (conditional) rank correlations assigned to the arcs of the NPBBN.

We are interested in *what if?* scenarios, in diagnosis and/or prediction, and moreover in visualisations and comparisons with the default situation. Examine the situation in which there is a very high risk of cancer. To do that, 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. In this case the NPBBN is used for diagnosis.

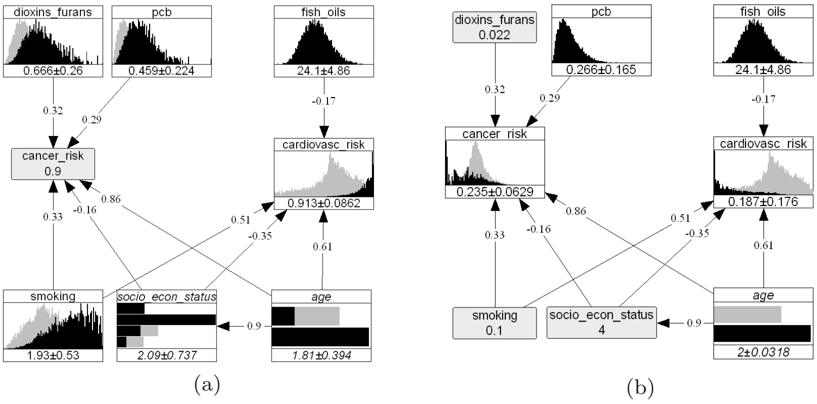

Fig. 1.13. Diagnostic & predictive reasoning using the NPBBN. (a) Conditionalised NPBBN for cancer_risk = 0.9. (b) Conditionalised NPBBN for dioxins_furans=0.022, smoking=0.1, socio_econ_status=4.

Fig. 1.12 and 1.13 are obtained with UniNet, a software application where the approach to mixed non-parametric continuous & discrete BBNs has been implemented[k]. In Fig. 1.13 the grey distributions in the background are the unconditional marginal distributions, provided for comparison. The conditional means and standard deviations are displayed under the histograms. In Fig. 1.13(a) we examine the situation of a very high cancer risk. We are interested in what can we infer about the factors influencing the cancer risk, when this is known to be 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 to have a high cancer risk. Because some of this factors influence also the cardiovascular risk, the shift in their distributions causes an increase in the cardiovascular risk as well.

The conditionalisation in a NPBBN can also be 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/furans. Fig. 1.13(b) presents the flow of this information through the graph. 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.

All the results and computations performed in this section are also possible if the model used is a regular vine rather then a NPBBN. Nevertheless the visualisation of such results is not yet available and the interpretations, in terms of the flow of influences, might be somewhat cumbersome when using regular vines.

One might wonder how we actually calculated the conditional distributions presented in Fig. 1.13. There are several ways to perform conditionalisation in NPBBNs.

Since sampling a NPBBN is based on the sampling procedure for regular vines, the cumulative or density approach for vines, mentioned in the beginning of this section, can be used to perform inference in NPBBNs. Whichever of the two methods is preferred, if the DAG contains undirected cycles, multiple integrals need to be evaluated for each sample, and for any new conditionalisation. This might be a very time consuming operation. Nevertheless, the problem owner might not be prepared to wait days or

[k] The software is available on http://dutiosc.twi.tudelft.nl/∼risk/, together with supporting scientific documentation.

not even hours for the results of new scenarios and policies. In these cases the advantages of fast updating algorithms for discrete BBNs [3, 16] are decisive. The reduced assessment burden and modelling flexibility of the NPBBNs are combined with the fast updating algorithms of discrete BBNs in the hybrid method presented in [5]. Sampling a large NPBBN structure once, and then discretizing it so as to enable fast updating provides an elegant solution to the above problem. This method is not applicable when working with regular vines, since no fast algorithms for vines on discrete variables are available.

The last and fastest way of conditionalising in a NPBBN is in the particular case in which the normal copula is used to realise the rank correlations. Since all the calculations are performed on a joint normal vine, any conditional distribution will also be normal, hence in this case conditioning can be performed analytically. This last method is implemented in UniNet, hence it was used to produce Fig. 1.13.

The advantages of the normal copula are also used in the next section where the model learning problem is discussed.

1.3. Data Mining with NPBBNs

In situations when data does not exist or is very limited, expert judgement must be used to define the graphical structure and assess the required parameters. However, if the data are available we would like to extract a fitting model from data. In the process of learning a model from data, two aspects can be of interest: learning the parameters of the model, given the structure, and learning the structure itself. Both learning the parameters of a regular vine, given the structure, and learning the vine structure together with its parameters are discussed in Chapter 3. The ensuing discussion concentrates on learning the DAG of a NPBBN together with its parameters from an ordinal data set.

The idea behind model inference for NPBBNs coincides with the one for regular vines, and it is based on the factorisation of the determinant of the correlation matrix on the arcs of the NPBBN. This factorisation is similar with the one for regular vines and the proof of this fact is available in [7]. Once again, the directed nature of a NPBBN and the possibility of excluding arcs that correspond to zero rank correlations made learning a NPBBN a more intuitive task than learning a regular vine.

A NPBBN induced from data can be used to investigate distant relationships between variables, as well as making predictions, by computing the conditional probability distribution of one variable given the values of some others (see the previous section).

The distinctive feature of learning a NPBBN from a data set is that the one dimensional marginal distributions are taken directly from data, and the model assumes only that the joint distribution has a normal copula. That is to say that the variables' rank dependence structure is that of a joint normal distribution. The NPBBN methodology is based on representing (conditional) dependencies on the arcs of a DAG, hence our strategy for inferring a NPBBN from data searches conditional dependencies in the data and associates arcs to them. A detailed discussion is found in [7]; here we only sketch the ideas.

The concepts of learning and validation are closely connected, as indeed the goal is to learn a NPBBN that is valid. Validation involves two steps: validating that the joint normal copula adequately represents the multivariate data, and validating that the NPBBN is an adequate model of the saturated graph. Validation requires an overall measure of multivariate dependence on which statistical tests can be based. A suitable measure in this case is the determinant of the rank correlation matrix [7]. The determinant is 1 if all variables are independent, and 0 if there is linear dependence between the normal versions of the variables. We distinguish 3 determinants: DER is the determinant of the empirical rank correlation matrix. DNR is the determinant of the rank correlation matrix obtained by transforming the marginals to standard normals, and then transforming the product moment correlations to rank correlations using Pearson's transformation[1]. Finally DBBN is the determinant of the rank correlation matrix of a NPBBN using the normal copula. DNR will generally differ from DER because DNR assumes the normal copula, which may differ from the empirical copula. A 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 NPBBN which represents the DNR parsimoniously. The saturated NPBBN will induce a joint distribution whose rank determinant is equal to DNR, since the NPBBN uses the normal copula. However, many of the influences only reflect sample jitter and we will eliminate them from the model. Moreover, for a large number of variables, the saturated graph is dense and unintu-

[1] Pearson's transformation[17] is characteristic to the normal distribution. The normal copula assumption implies that the variables are assumed to have the distribution of transforms of a joint normal vector.

itive.

Once the normal copula is validated we will build the NPBBN by adding arcs between variables only if the rank correlation between those two variables is among the largest. The second validation step is similar to the first. The general procedure can then be represented thus:

(1) Verify that DER is not outside the plausible central confidence band for DNR. If so, the normal copula hypothesis is not rejected;
(2) Construct a skeletal NPBBN by adding arcs to capture known causal or temporal relations;
(3) If DNR is within the 90% central confidence band of the determinant of the skeletal NPBBN, then stop, else continue with the following steps;
(4) Find the pair of variables such that the arc between them is not in the DAG and their rank correlation is greater than the rank correlation of any other pair not in the DAG. Add an arc between them 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 procedure for building a NPBBN to represent a given data set is not fully automated, as it is impossible to infer directionality of influence from multivariate data. Insight into the causal processes generating the data should be used, whenever possible, in constructing a NPBBN. Because of this fact, there are different NPBBN structures that are wholly equivalent, and many non-equivalent NPBBNs may provide statistically acceptable models of a given multivariate ordinal data set.

This approach is already used in several studies that try to link $PM_{2.5}$ concentrations to stationary source emissions [7, 8, 15]. Other applications of the NPBBN methodology are briefly mentioned in the next section.

1.4. Applications of NPBBNs

In Example 1.2 we have already mentioned one of the ongoing applications that uses NPBBNs, namely *Beneris*. *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[10].

Another project which uses NPBBNs is *CATS*, which 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[1, 2].

It is worth mentioning that both *Beneris* and *CATS* models use NPBBNs with hundreds of nodes and arcs. Models involving hundreds of variables benefit greatly form the advantages of the directed structure of a NPBBN. The use of regular vines in such situations would be somewhat cumbersome if not impossible.

A third application employs NPBBNs as a tool to estimate the extent of a fire in a building, given any combination of possible conditions and any unexpected course of events during an emergency[9].

The latest attempt to use a NPBBN based approach is in the field of reservoir engineering, namely in the estimation of surface characteristics (see www.data-assimilation.com/ssda).

All of the above projects use UniNet, the software application mentioned in Section 1.2.2. UniNet was initially developed to support the *CATS* project, and it is under constant development. The main program features are presented in the Appendix of [4].

1.5. Conclusions

In the present book graphical models have been chosen to represent multivariate distributions with complex dependence structures. More specifically, regular vines were advocated for this purpose. This chapter proposes NPBBNs as an alternative to regular vines and discusses the differences and similarities between the two.

The most important difference between NPBBNs and regular vines turned out to be the different statements of conditional (in)dependence that they make through their undirected and directed nature, respectively. In the DAG of a NPBBN the absence of an arc encodes (conditional) independence statements. Regular vines on the other hand can be viewed as fully connected graphs that represent (conditional) dependence statements. Accordingly the *absence* of edges in a regular vine is only possible for very special structures[m]. Nevertheless, the presence of arcs in NPBBNs does not guarantee dependence between variables (see Remark 1.1). Consequently if one graph fails to represent dependencies, the other one fails to represent independencies.

The possibility of excluding arcs from a NPBBN, whenever a (condi-

[m] If all conditional rank correlations in the higher order trees of a vine are zero, then the edges of these trees can be removed[12].

tional) independence statement is known, has the advantage of a sparser resultant graphical structure, which is often more readable. In order to *visualise* (conditional) independence statements on a regular vine one has to resort to assigning zero (conditional) rank correlations to the edges. In this way, similar independence statements can be represented using both structures, and comparisons can be made. After such an analysis, no definite conclusion emerged. Some combinations of statements are better represented using a regular vine, whereas others benefit from the representation in a DAG form. Nevertheless, this is only true for small structures. When hundreds of variables are involved, the saturated nature of regular vines constitutes a great disadvantage in modelling and visualising. Moreover the directed structure of NPBBNs holds the advantage of a more intuitive representation in terms of the flow of influences between variables.

When it comes to the quantitative part of the models, both NPBBNs and regular vines require marginal distributions and (conditional) rank correlations. Once these are obtained, the joint distribution is stipulated through a sampling procedure. The sampling procedure for NPBBNs uses the one for regular vines, hence we cannot talk about the advantages of the former compared to the latter. Moreover in DAG structures, that contain large undirected cycles, sampling a NPBBN involves extra numerical calculations that might be time consuming. These calculations are not necessary if the multivariate distribution can be represented and assessed using a regular vine. However, this disadvantage of NPBBNs vanishes when the normal copula is used.

Possessing a joint distribution allows us to perform inference. We can calculate the conditional distributions of unobserved variables, given the values of the observed ones. To achieve this, similar calculations are performed in both graphical models. Numerical complications that might arise for DAGs containing undirected cycles are circumvented by using a hybrid method that combines the flexibility of NPBBNs with the fast updating algorithms of discrete BBNs. When regular vines are used, the new conditional distributions, although calculated, cannot be easily visualised and compared with the unconditional one. This is purely an implementation issue for graphical software, hence it might be viewed as a recommendation for future development. Nonetheless NPBBNs hold the advantage that conditionalisation can be interpreted in terms of the directionality of arcs. In other words, if the reasoning is done ”bottom-up” (in terms of the directionality) then the NPBBN is used for diagnosis, whereas if it is done ”top-down”, the NPBBN serves for prediction.

When data are available we are interested in learning a fitting model from data. In this process we could either learn the parameters of the model, given the structure, or learn the structure itself. The subject of learning the parameters of a NPBBN given the structure was not yet addressed. Future research could investigate the methodology presented in Chapter 3 and its applicability to NPBBNs.

The idea behind learning the DAG of a NPBBN together with its parameters from an ordinal data set coincides with the one for learning regular vines. Still, the directed nature of a NPBBN and the possibility of including only arcs that correspond to the highest rank correlations make learning a NPBBN a more intuitive task than learning a regular vine.

References

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

[2] Ale, B., Bellamy, L.J., R. van der Boom, Cooper, J., Cooke, R., Goossens, L.H.J., Hale, A.R., Kurowicka, D., Morales, O., Roelen, A. and Spouge, J. (2009), Further development of a causal model for air transport safety (cats); building the mathematical heart. *Reliability Engineering and System Safety Journal.*

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

[4] Hanea, A. (2008). *Algorithms for Non-Parametric Bayesian Belief Nets*, PhD Dissertation, Delft Institute of Applied Mathematics (Wöhrmann Print Service).

[5] Hanea, A., Kurowicka, D. and Cooke, R. (2006). Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets, *Quality and Reliability Engineering International* **22**, 6, pp. 613–729.

[6] Hanea, A., Kurowicka, D. and Cooke, R. (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.*

[7] Hanea, A., Kurowicka, D., Cooke, R. and Ababei, D. (2007). Mining and Visualising Ordinal Data with Non-Parametric Continuous BBNs, *Computational Statistics and Data Analysis*, **10.1016/j.csda.2008.09.032.**.

[8] Hanea, A. M. and Harrington, W. (2009). Ordinal $PM_{2.5}$ Data Mining with Non-Parametric Continuous Bayesian Belief Nets.

[9] Hanea, D. and Ale, B. (2009). Risk of human fatality in build-

ing fires: A decision using Bayesian networks, *Fire Safety Journal* **10.1016/j.firesaf.2009.01.006**.

[10] Jesionek, P. and Cooke, R. (2007). Generalized method for modeling dose-response relations application to BENERIS project, Tech. rep., European Union project.

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

[12] Kurowicka, D. and Cooke, R. (2006). Completion problem with partial correlation vines. *Linear Algebra and Its Applications* **418(1)**, pp. 188–200.

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

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

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

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

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

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

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