Mining and Visualising Ordinal Data with Non-Parametric Continuous BBNs

A.M.Hanea* D.Kurowicka R.M.Cooke D.A.Ababei
Institute of Applied Mathematics
Delft University of Technology
The Netherlands

Abstract

Data mining is the process of extracting and analysing information from large databases. Graphical models are a suitable framework for probabilistic modelling. A Bayesian Belief Net(BBN) is a probabilistic graphical model, which represents joint distributions in an intuitive and efficient way. It encodes the probability density (or mass) function of a set of variables by specifying a number of conditional independence statements in the form of a directed acyclic graph. Specifying the structure of the model is one of the most important design choices in graphical modelling. Notwithstanding their potential, there is only a limited number of applications of graphical models on very complex and large databases. A method for mining ordinal multivariate data using non-parametric BBNs is presented. The main advantage of this method is that it can handle a large number of continuous variables, without making any assumptions about their marginal distributions, in a very fast manner. Once the BBN is learned from data, it can be further used for prediction. This approach allows for rapid conditionalisation, which is a very important feature of a BBN from a user's standpoint.

Keywords: ordinal data mining, non-parametric continuous bayesian belief nets, vines, copula, structure learning.

*Correspondence to: A.M.Hanea, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands, Telephone: +31 1527 84563, Fax: +31 1527 87255, E-mail: A.Hanea@ewi.tudelft.nl.

1 Introduction

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. We describe a method for mining ordinal multivariate data using non-parametric Bayesian Belief Nets (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 database contains monthly emissions of SO_2 and NO_x at different locations, and monthly means of the readings of $PM_{2.5}$ concentrations at various monitoring sites. SO_2 is the formula for the chemical compound sulfur dioxide. This gas is the main product from the combustion of sulfur compounds and is of significant environmental concern. 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. The notation $PM_{2.5}$ is used to describe particles of 2.5 micrometers or less in diameter.

There are 786 emission stations and 801 collection sites. For most emission 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. 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 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.

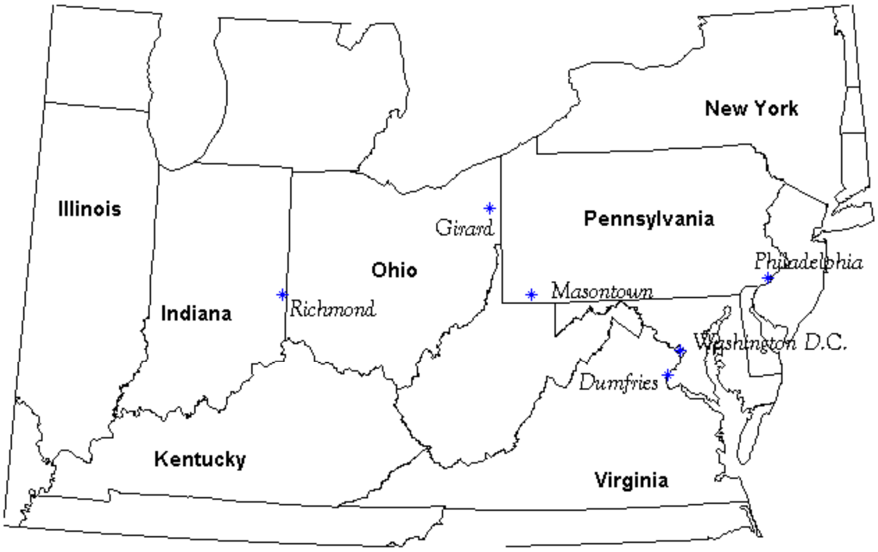


Figure 1: Selected power plant emissions.

Definitions and concepts are introduced in Section 2, but suffice to say now that BBNs are directed acyclic graphs where an arrow connecting a parent node to a child node indicates that influence flows from parent to child. A BBN for Washington DC ambient $PM_{2.5}$ is shown in Figure 2. This model is similar to the one described and analysed in [25]. 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); 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 (see [24]), but a smaller number of variables is more suitable for explaining the method.

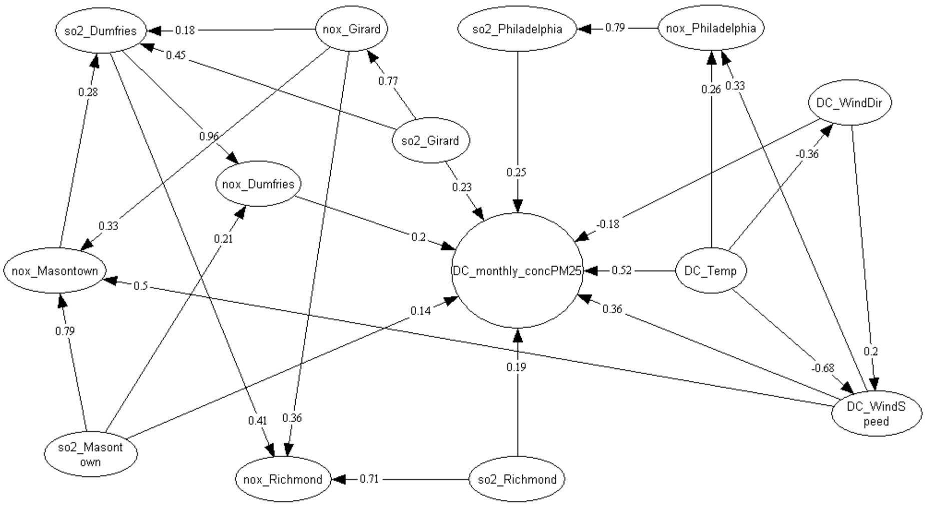


Figure 2: BBN for Washington DC ambient $PM_{2.5}$.

Most BBNs are discrete and arrows represent mathematical relationships in form of conditional probability tables. If the number of possible values for nodes is modestly large (in the order 10) such models quickly become intractable. Thus, a variable like DC_monthly_concPM25 (see Figure 2) with all variables discretised to 10 possible values, would require a conditional probability table with 10^9 entries. So-called discrete continuous BBNs [6] allow continuous nodes with either continuous or discrete parents, but they assume that the continuous nodes are joint normal. Influence between continuous nodes is represented as partial regression coefficients [27; 33]. The restriction to joint normality is rather severe. Figure 3 shows the same BBN as Figure 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

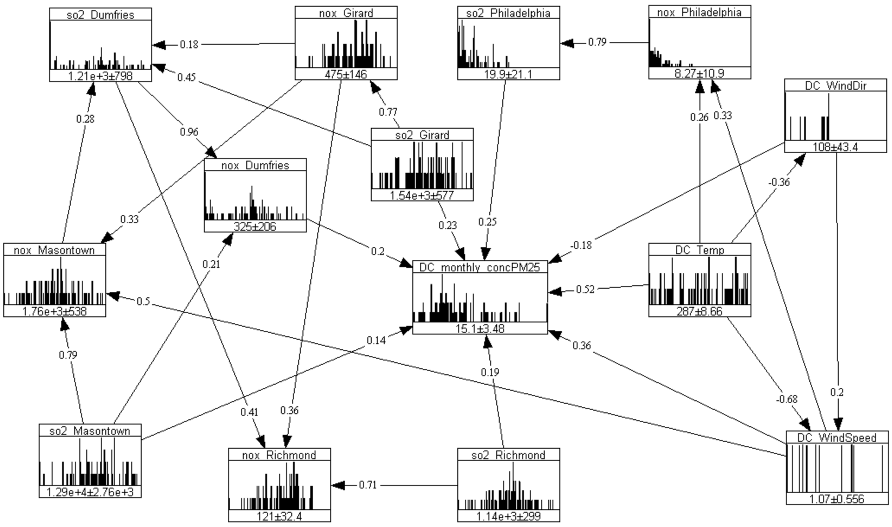


Figure 3: Washington DC ambient $PM_{2.5}$ BBN with histograms.

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) Spearman's rank correlations according to a protocol explained in Section 2. 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. Conditionalisation is performed on the transformed variables, which are assumed to follow a joint normal distribution, hence any conditional distribution will also be normal with known mean and variance. Finding the conditional distribution of a 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 [8].

Rapid conditionalisaion is a very important feature of a BBN from a user's standpoint. To illustrate, Figures 4 and 5 show the result of conditionalising the joint distribution on cold weather (275K) in Washington and low (Figure 4) and high (Figure 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 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 3.2 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 fine particulate 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 Pearson's transformation (see Section 2). 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 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 straightforward 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 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 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.

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

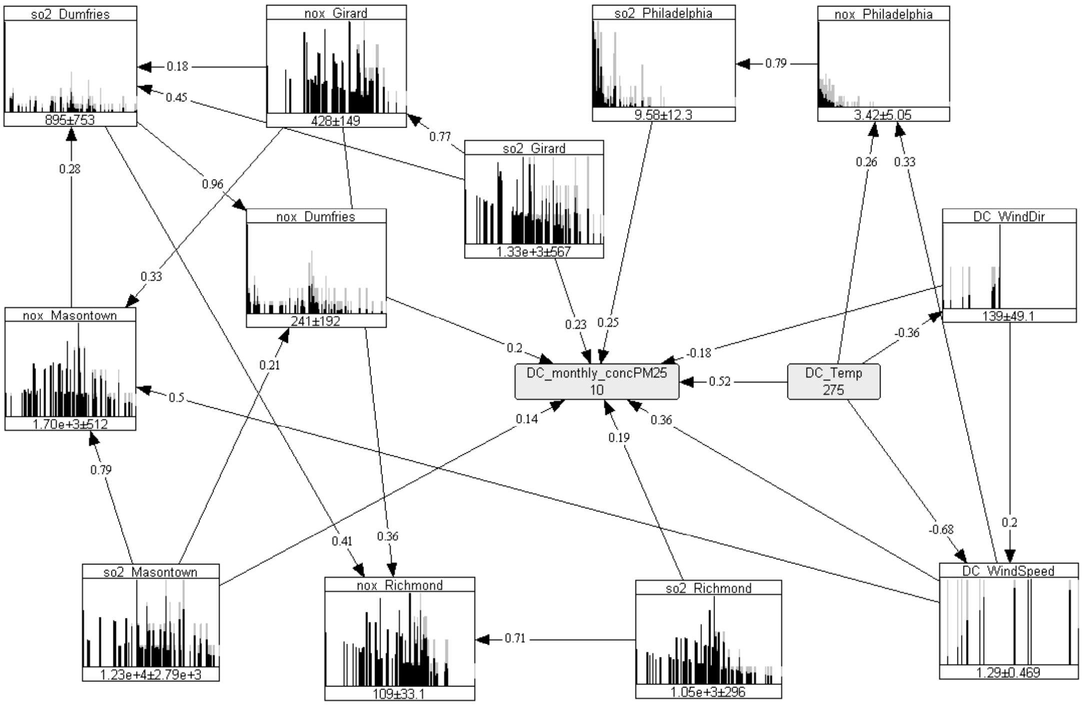


Figure 4: Conditionalisation on low concentration of $PM_{2.5}$ for Washington DC and cold weather.

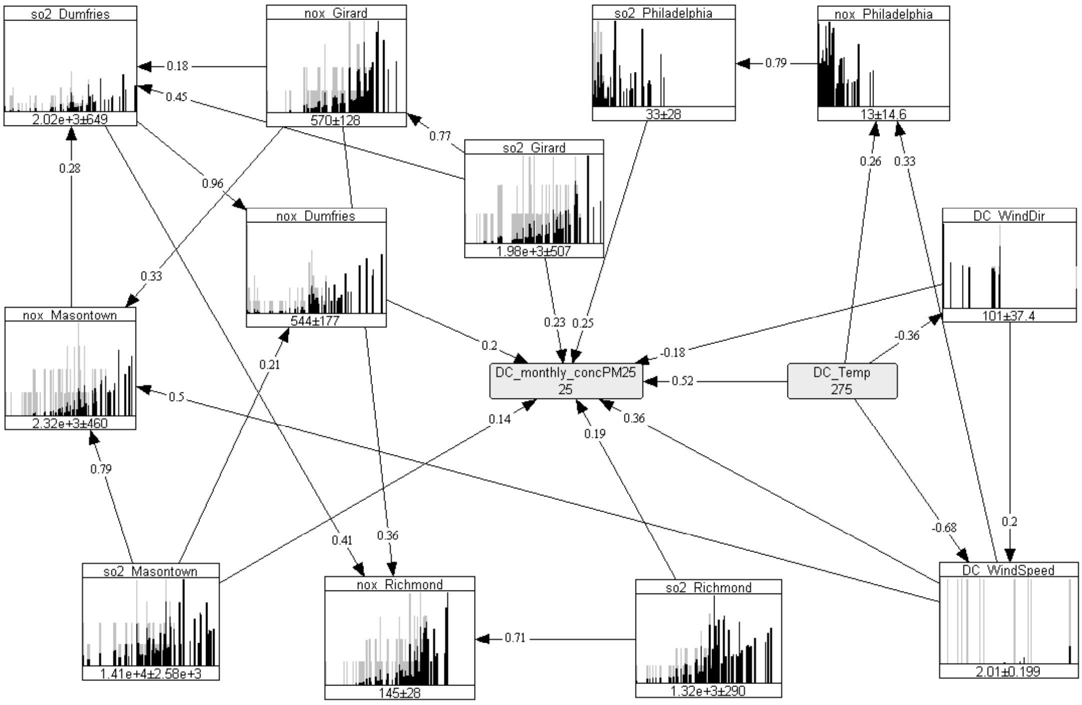


Figure 5: Conditionalisation on high concentration of $PM_{2.5}$ for Washington DC and cold weather.

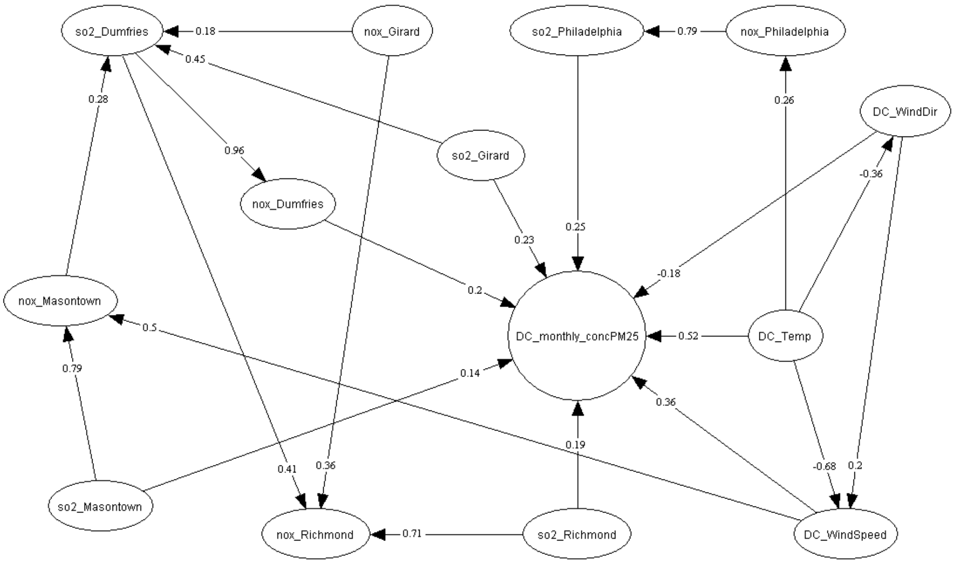


Figure 6: Simplified BBN with 22 arcs.

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 [8] and are reviewed in Section 2. The key notion is to link a BBN with a nested sequence of regular vines.

In Section 3.1 we present a short overview of the existing methods for learning the structure of a BBN from data. In order to introduce our approach we need to select a measure of multivariate dependence. Section 3.2 contains a discussion of various such measures. In Section 3.3 we introduce our learning algorithm, and in Section 4 we present this approach using the database of pollutants emissions and fine particulate concentrations. In the last part of this paper we discuss alternative ways to calculate the correlation matrix of a BBN and illustrate how these may speed up the updating algorithm.

2 Definitions & Preliminaries

In this section we present, in a more formal fashion, concepts that are used in learning the structure of a BBN. We discuss non-parametric continuous BBNs and their relationship with the graphical model vines.

A BBN encodes the probability density (or mass) function of a set of variables by specifying a number of conditional independence statements in a form of a directed acyclic graph and a set of conditional distribution functions of each variable given its parents in the graph. In Figure 2 we see that the variable so2_Richmond does not have any parents but is a parent of nox_Richmond, and of DC_monthly_concPM25. nox_Richmond (or DC_monthly_concPM25) is called a child of so2_Richmond. In non-parametric continuous BBNs nodes represent continuous variables with invertible distribution functions and arcs are associated with (conditional) rank correlations. Therefore, every arc in the BBN is assigned a (conditional) rank correlation between parent and child. For example, in Figure 2 the arc between nox_Philadelphia and its parent DC_Temp is associated with the rank correlation between these variables equal to 0.26. The arc between nox_Philadelphia and its other parent DC_WindSpeed is associated with the rank correlation between these variables, conditional on DC_Temp, the parent with which the correlation has been already assigned (for details see [18]). In general, if we denote with i the node nox_Philadelphia, and with j, DC_WindSpeed, then DC_Temp is denoted D_{ij} and it represents the conditioning set for the arc between nodes i and j. The value of this conditional rank correlation is 0.33. An important result is that the assignment of (conditional) rank correlations to the arcs of a BBN is unconstrained, meaning that we can assign to these arcs any number between -1 and 1 and each such assignment will be consistent. This property follows from a close relationship between non-parametric continuous BBNs with vines.

Vines were introduced in [4; 1]. 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, and each tree has the maximum number of edges. A *regular vine* on n variables ia a vine on which two edges in tree j are joined by an edge in tree $j+1$ only if these edges share a common node. Further in this paper, whenever we speak of vines, we mean regular vines.

Figure 7 shows a vine that represents the same joint distribution as the BBN in Figure 2. We replaced the name of each variable by its number in a sampling order. The numbers from 14 to 1 correspond to DC_monthly_concPM25, nox_Dumfries, nox_Richmond, so2_Philadelphia, so2_Dumfries, nox_Philadelphia, nox_Masontown, DC_WindSpeed,DC_WindDir, nox_Girard, so2_Richmond, DC_Temp, so2_Girard, so2_Masontown, respectively. 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. For example, the edges $(10, 9)$ and $(9, 8)$ of the vine from Figure 7, are joined by an edge in the second tree. The conditioned set of this edge is $(10, 8)$ and the conditioning set is (9).

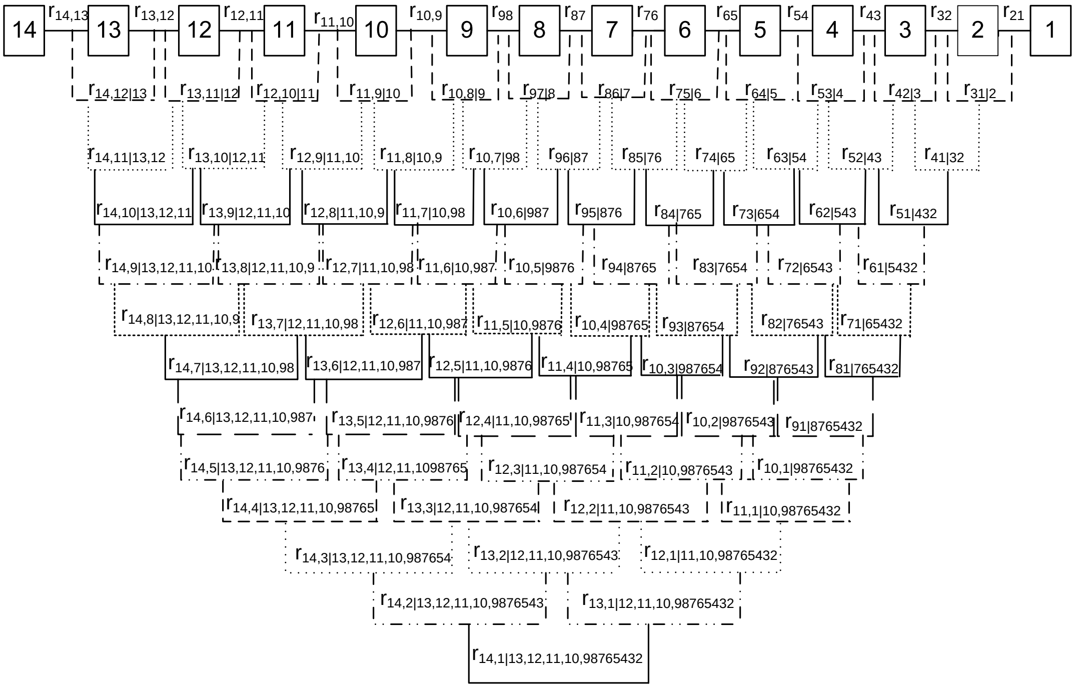


Figure 7: Vine for Washington DC ambient $PM_{2.5}$.

Each edge of this regular vine is associated with a (conditional) rank correlation, denoted by r, just like the arcs of the BBN. These (conditional) rank correlations can be arbitrarily chosen in the interval $[-1, 1]$. Using a copula to realise them, a joint distribution satisfying the copula-vine specification can be constructed and it will always be consistent.

The joint distribution of a set of variables can be graphically represented as a BBN or as a regular 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 edges. In some cases these two structures require exactly the same (conditional) rank correlations. But 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. 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. In the process of translating a rank correlation specification for a BBN into a rank correlation specification for a vine 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 [8], as we can always recalculate required correlations for a given ordering of the variables. This is due to a number of properties of the normal copula that are discussed below.

Some important properties of vines translate almost immediately to corresponding properties of non-parametric BBNs. We will further present these properties.

Each vine edge may be associated with a partial correlation. The result is a partial correlation vine specification. Partial correlations are defined as geometric averages of partial regression coefficients [36]. They can be calculated from correlations in the correlation matrix using the following recursive formula for partial correlation [18]:

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

In [1] it is shown that each such partial correlation vine specification uniquely determines the correlation matrix, and every full rank correlation matrix can be obtained in this way.

A partial correlation vine specification does not uniquely specify a joint distribution, moreover a given set of marginal distributions may not be consistent with a given set of partial correlations. Nevertheless there is a joint distribution satisfying the specified information [1]. For example a joint normal distribution. 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. In general, conditional independence is neither necessary, nor sufficient for zero partial correlation [17]. This property of the joint normal variables follows from two facts: 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 Pearson's transformation [29]: $\rho(X,Y) = 2\sin(\frac{\pi}{6} \cdot r(X,Y))$, and it translates these properties to normal copula.

Vines are actually a way of factorising the determinant of the correlation matrix. For any vine on n variables, the product of one minus squared partial correlations assigned to the edges of the vine, is the same, and equal to the determinant of the correlation matrix [18]:

$$D = \prod_{e \in E(\mathcal{V})} \left(1 - \rho^2_{e_1,e_2;D_e}\right), \tag{2.2}$$

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.

The key notion in the derivation of equation 2.2 is the *multiple correlation* and its relationship with partial correlations. The multiple correlation $R_{1:2,\ldots,n}$ of variable 1 with respect to $2,\ldots,n$ 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. It is the correlation between 1 and the best linear predictor of 1 based on $2,\ldots,n$. It is easy to show that [18]:

$$D = \left(1 - R^2_{1:2,\ldots,n}\right)\left(1 - R^2_{2:3,\ldots,n}\right)\ldots\left(1 - R^2_{n-1:n}\right). \tag{2.3}$$

In [16] it is shown that $R_{1:2,\ldots,n}$ is non negative and satisfies:

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

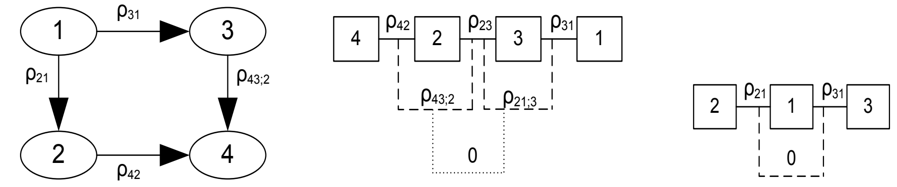
A similar factorisation of the determinant of the correlation matrix holds for the partial correlation specification for BBNs. This factorisation plays a central role in model inference from Section 3.

Theorem 2.1. *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 nodes i and j, with conditioning set D_{ij}, 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 equation 2.2. 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 equation 2.2 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 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 8a. A sampling order of these variables is 1,2,3,4. The vine corresponding to this BBN is shown in Figure 8b.



(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 8: 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 equation 2.2. 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 8c. 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 above concepts and results will be used in our learning algorithm. A partial correlation BBN fully characterises the correlation structure of the joint distribution and the values of the partial correlations are algebraically independent. Unlike the correlations in a correlation matrix, the partial correlations in a BBN need not satisfy an algebraic constraint like positive definiteness. Moreover, the partial correlation BBN represents a factorisation of the determinant of the correlation matrix. The determinant of the correlation matrix is a measure of linear dependence in a joint distribution. If all variables are independent, the determinant is 1, and if there is linear dependence between the variables, the determinant is 0. Intermediate values reflect intermediate dependence. Our learning algorithm will choose a structure for which only arcs corresponding to large partial correlations are present. Thus we will remove arcs that correspond to small partial correlations. Hence, we will change small partial correlations to zero while disturbing the determinant as little as possible.

3 Learning the Structure of a BBN

3.1 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 [6].

In the process of learning a BBN from data, two aspects are of interest: learning the parameters of the BBN, given the structure, and learning the structure itself. We focus on structure learning. A vast literature is available on this subject. Neither the space, nor the purposes of this article permits a complete overview of existing learning methods. For example, we will omit from our discussion naive BBNs.

Most of the current methods to learn the structure of a BBN focus on discrete or gaussian variables [34]. 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 [20; 10].

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 [2; 34; 28].

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 [5]; [12]; [2]). Unfortunately, this information is not always available.

To our knowledge, the few methods that can handle non-parametric continuous variables (e.g.,[22]) 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 [3] 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 distribution of the variables. We want to be able to learn such structures fast and use them further, for prediction purposes.

The learning algorithm presented in Section 3.3 has been implemented into a software application, called UniNet. UniNet allows for quantification of mixed discrete & non-parametric continuous BBNs (the theory for non-parametric continuous BBNs is extended in [9] to include ordinal variables). The program has a friendly interface and the simulations are very fast. BBNs with thousands of nodes can be conditionalized on arbitrary values of random variables, whereby the conditional distribution is computed in a few minutes. Moreover, UniNet will shortly be (freely) available on the Internet.

Comparisons of our method for learning the structure of a non-parametric continuous BBN with other existing methods are difficult to conduct. There are several reasons for that. 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. [11]; [15]), and the other one is to use the discretised version of the variables (e.g. [7]). We use neither of the two methods.

An algorithm that deals with non-parametric continuous variables is proposed in [22]. The authors of [22] 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 [34]. 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 [22], 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 proves to us that the method is inadequate for the problem so it would not be useful to compare it with ours.

3.2 Multivariate Dependence Measures

Inferring the structure of a BBN from data requires a suitable measure of multivariate dependence. Multivariate dependence measures are discussed in [23; 14; 32]. In [23] Renyi's axioms [31] 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 [23].

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 a 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 [23], and AX 7.1 & AX 7.2 are a bit weaker than their counterpart in [23]. Axioms 6.1 and 6.2 explicitly capture the notion of linear dependence.

Proposition 3.1. *$D_{1,...,n} = Det(C)$, 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:

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

To prove that axioms 6.1 and 6.2 hold, we use equation 2.3. $D_{1,...,n}$ is zero if and only if at lest one of the multiple correlations in equation 2.3 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 3.1. *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 [13] 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 [32] is sometimes proposed as a measure of multivariate dependence. For bivariate dependence, the Spearman's rank correlation is given by [26]:

$$r(X_1, X_2) = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3 = 12 \int_0^1 \int_0^1 uvc(u, v) du dv - 3. \tag{3.1}$$

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 3.1 are not equal and a variety of possible generalisations exist [32]. In three dimensions the version based on the copula density reads [32]:

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

Using the bivariate elliptical copula [19] 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 $E(U|V = v) = v\rho(U, V)$, [18], 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.

3.3 Learning the Structure of a Non-Parametric 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 Section 2);
- DBBN = the determinant of the rank correlation matrix of a BBN using the normal copula.

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 [21] 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 [35], 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 [35] 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 [35] 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 described in Section 2.

If the BBN is not saturated, then DBBN > DNR. We will use the result from Theorem 2.1 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 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 BBN model. Because of this fact, there are different BBNs that are wholly equivalent, and many non-equivalent BBNs may provide statistically acceptable models of a given multivariate ordinal data set.

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 correlation 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. This choice will only make a difference if the number of samples is very large.

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.

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

We illustrate our method for learning a BBN from data using an ordinal multivariate data set that we briefly introduced in Section 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 collection 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 9). The goal is to build a BBN that captures the dependence structure between these variables, using the approach presented in the previous section. All analysis and graphs are produced by the UniNet software.

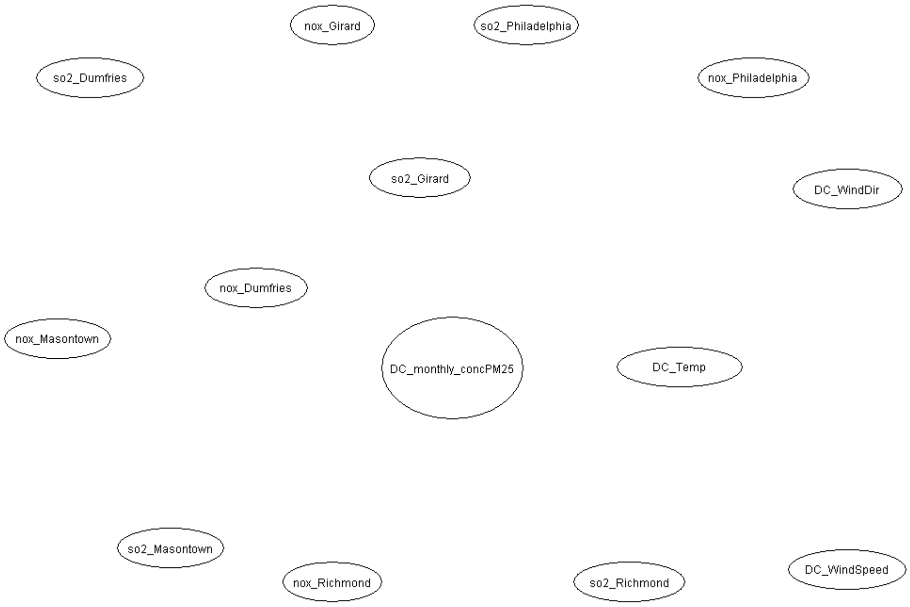


Figure 9: BBN on 14 nodes with no arcs.

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 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 10 and 11 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 [30]:

Table 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 9 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 11). 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 12. UniNet calculates from data the (conditional) rank correlations that correspond to the arcs of the BBN.

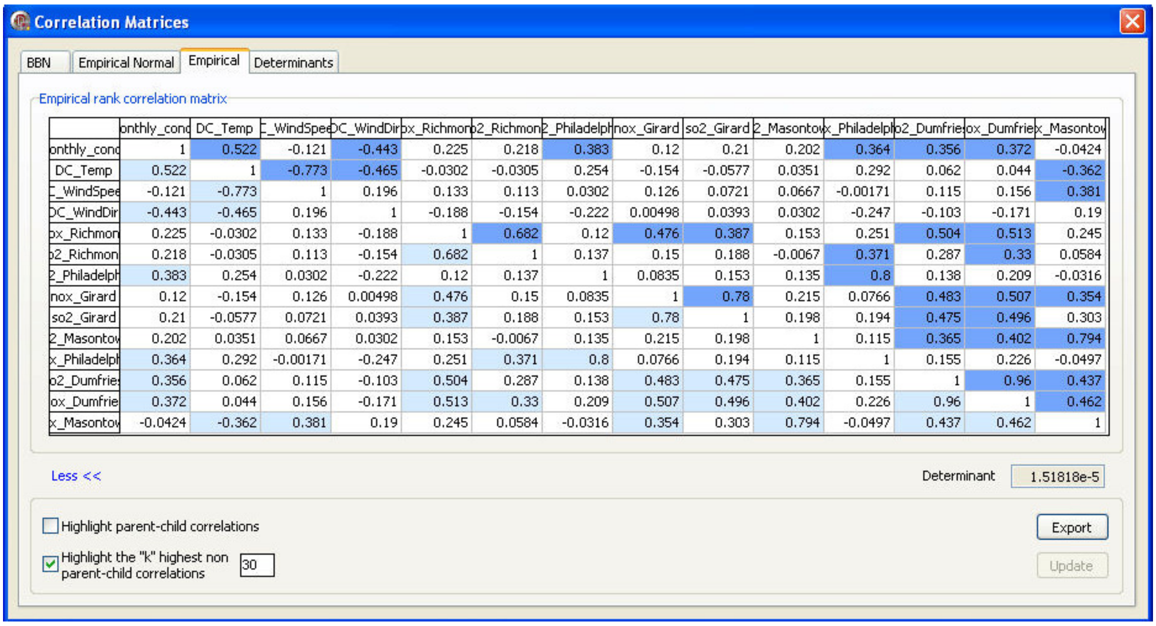
The determinant of the rank correlation matrix based on the new BBN differs from DNR, as this BBN is not saturated. It hypothesises conditional independence where the data exhibits small partial correlations. In this case DBBN = 3.6838E-04 and its 90% central confidence interval is $[0.5552\text{E-}04, 3.7000\text{E-}04]$. We notice that 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 13.

The 90% central confidence interval for the determinant of the rank correlation matrix based on the new BBN is $[0.4617\text{E-}04, 2.6188\text{E-}04]$ 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 14.

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



Empirical rank correlation matrix

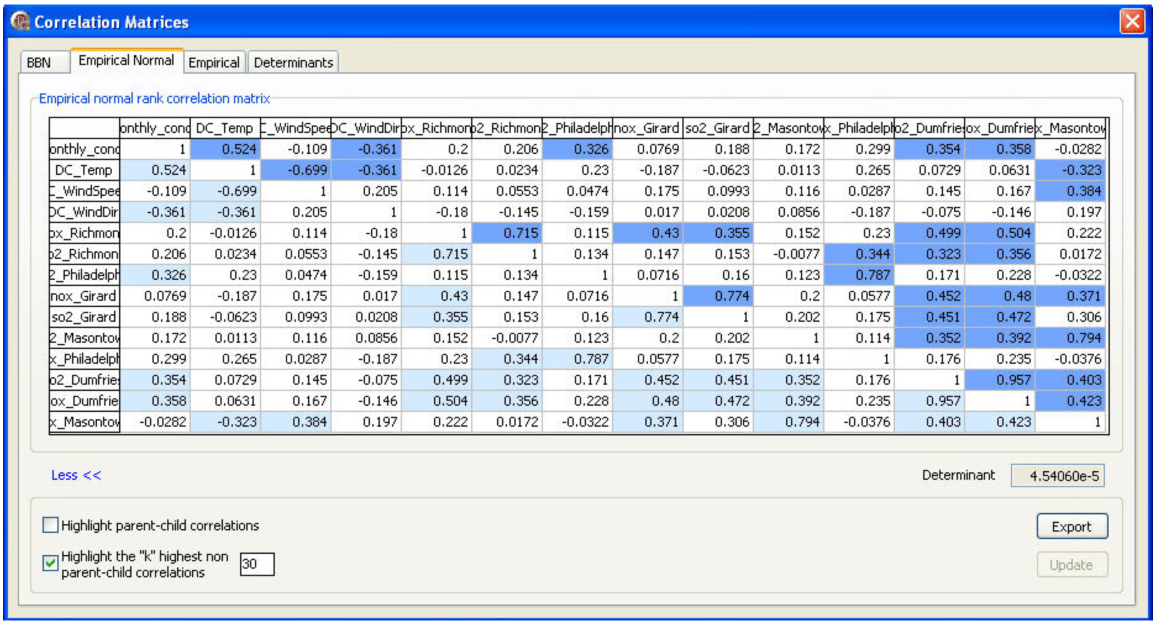
	onthly_cond	DC_Temp	C_WindSpee	DC_WindDir	ox_Richmon	o2_Richmon	2_Philadelph	nox_Girard	so2_Girard	2_Masontow	x_Philadelph	o2_Dumfries	ox_Dumfrie	x_Masontow
onthly_cond	1	0.522	-0.121	-0.443	0.225	0.218	0.383	0.12	0.21	0.202	0.364	0.356	0.372	-0.0424
DC_Temp	0.522	1	-0.773	-0.465	-0.0302	-0.0305	0.254	-0.154	-0.0577	0.0351	0.292	0.062	0.044	-0.362
C_WindSpee	-0.121	-0.773	1	0.196	0.133	0.113	0.0302	0.126	0.0721	0.0667	-0.00171	0.115	0.156	0.381
DC_WindDir	-0.443	-0.465	0.196	1	-0.188	-0.154	-0.222	0.00498	0.0393	0.0302	-0.247	-0.103	-0.171	0.19
ox_Richmon	0.225	-0.0302	0.133	-0.188	1	0.682	0.12	0.476	0.387	0.153	0.251	0.504	0.513	0.245
o2_Richmon	0.218	-0.0305	0.113	-0.154	0.682	1	0.137	0.15	0.188	-0.0067	0.371	0.287	0.33	0.0584
2_Philadelph	0.383	0.254	0.0302	-0.222	0.12	0.137	1	0.0835	0.153	0.135	0.8	0.138	0.209	-0.0316
nox_Girard	0.12	-0.154	0.126	0.00498	0.476	0.15	0.0835	1	0.78	0.215	0.0766	0.483	0.507	0.354
so2_Girard	0.21	-0.0577	0.0721	0.0393	0.387	0.188	0.153	0.78	1	0.198	0.194	0.475	0.496	0.303
2_Masontow	0.202	0.0351	0.0667	0.0302	0.153	-0.0067	0.135	0.215	0.198	1	0.115	0.365	0.402	0.794
x_Philadelph	0.364	0.292	-0.00171	-0.247	0.251	0.371	0.8	0.0766	0.194	0.115	1	0.155	0.226	-0.0497
o2_Dumfries	0.356	0.062	0.115	-0.103	0.504	0.287	0.138	0.483	0.475	0.365	0.155	1	0.96	0.437
ox_Dumfrie	0.372	0.044	0.156	-0.171	0.513	0.33	0.209	0.507	0.496	0.402	0.226	0.96	1	0.462
x_Masontow	-0.0424	-0.362	0.381	0.19	0.245	0.0584	-0.0316	0.354	0.303	0.794	-0.0497	0.437	0.462	1

Determinant 1.51818e-5

☐ Highlight parent-child correlations

☑ Highlight the "k" highest non parent-child correlations 30

Figure 10: The empirical rank correlation matrix of the 14-dimensional distribution.



Empirical normal rank correlation matrix

	onthly_cond	DC_Temp	C_WindSpee	DC_WindDir	ox_Richmon	o2_Richmon	2_Philadelph	nox_Girard	so2_Girard	2_Masontow	x_Philadelph	o2_Dumfries	ox_Dumfrie	x_Masontow
onthly_cond	1	0.524	-0.109	-0.361	0.2	0.206	0.326	0.0769	0.188	0.172	0.299	0.354	0.358	-0.0282
DC_Temp	0.524	1	-0.699	-0.361	-0.0126	0.0234	0.23	-0.187	-0.0623	0.0113	0.265	0.0729	0.0631	-0.323
C_WindSpee	-0.109	-0.699	1	0.205	0.114	0.0553	0.0474	0.175	0.0993	0.116	0.0287	0.145	0.167	0.384
DC_WindDir	-0.361	-0.361	0.205	1	-0.18	-0.145	-0.159	0.017	0.0208	0.0856	-0.187	-0.075	-0.146	0.197
ox_Richmon	0.2	-0.0126	0.114	-0.18	1	0.715	0.115	0.43	0.355	0.152	0.23	0.499	0.504	0.222
o2_Richmon	0.206	0.0234	0.0553	-0.145	0.715	1	0.134	0.147	0.153	-0.0077	0.344	0.323	0.356	0.0172
2_Philadelph	0.326	0.23	0.0474	-0.159	0.115	0.134	1	0.0716	0.16	0.123	0.787	0.171	0.228	-0.0322
nox_Girard	0.0769	-0.187	0.175	0.017	0.43	0.147	0.0716	1	0.774	0.2	0.0577	0.452	0.48	0.371
so2_Girard	0.188	-0.0623	0.0993	0.0208	0.355	0.153	0.16	0.774	1	0.202	0.175	0.451	0.472	0.306
2_Masontow	0.172	0.0113	0.116	0.0856	0.152	-0.0077	0.123	0.2	0.202	1	0.114	0.352	0.392	0.794
x_Philadelph	0.299	0.265	0.0287	-0.187	0.23	0.344	0.787	0.0577	0.175	0.114	1	0.176	0.235	-0.0376
o2_Dumfries	0.354	0.0729	0.145	-0.075	0.499	0.323	0.171	0.452	0.451	0.352	0.176	1	0.957	0.403
ox_Dumfrie	0.358	0.0631	0.167	-0.146	0.504	0.356	0.228	0.48	0.472	0.392	0.235	0.957	1	0.423
x_Masontow	-0.0282	-0.323	0.384	0.197	0.222	0.0172	-0.0322	0.371	0.306	0.794	-0.0376	0.403	0.423	1

Determinant 4.54060e-5

☐ Highlight parent-child correlations

☑ Highlight the "k" highest non parent-child correlations 30

Figure 11: The normal rank correlation matrix of the 14-dimensional distribution.

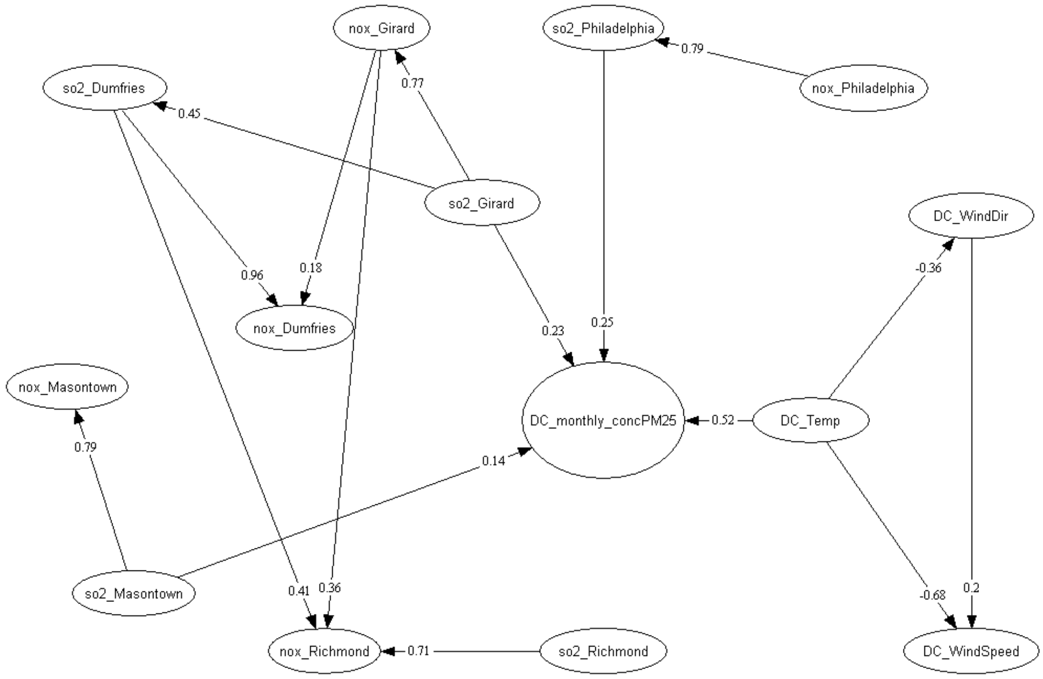


Figure 12: BBN on 14 nodes with 16 arcs.

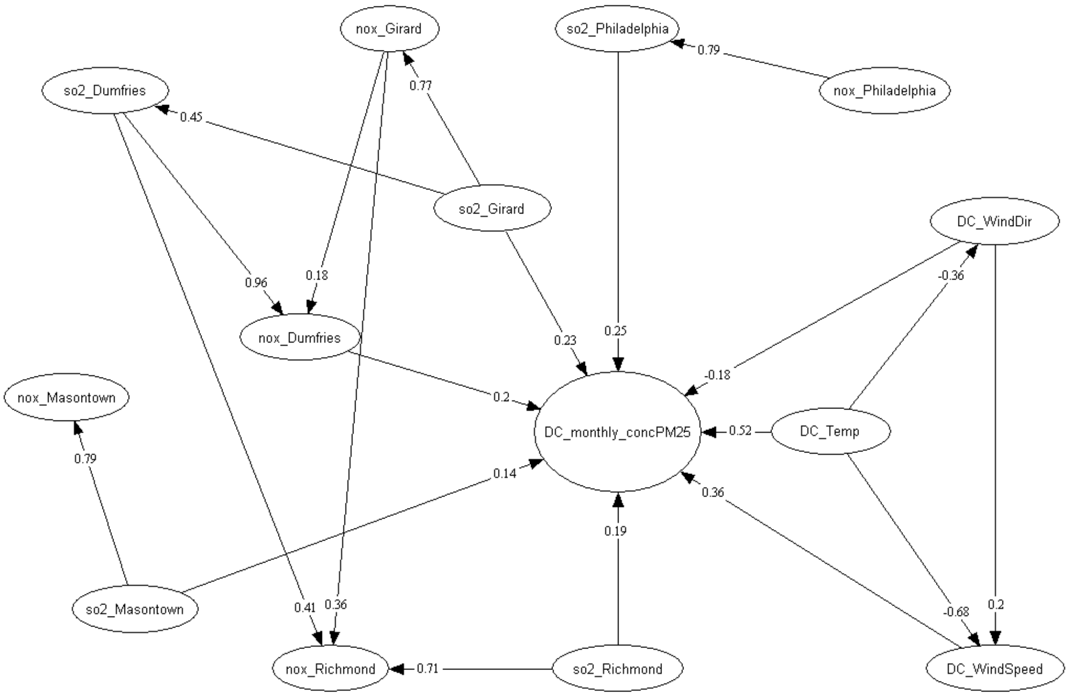


Figure 13: BBN on 14 nodes with 20 arcs.

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 15). 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 6, whose determinant is 4.8522E-04, and would be rejected. Using the 95% instead of the 90% central confidence interval would not change the conclusions of our analysis.

Another procedure for building a BBN to represent a given data set, would be to begin with one of the possible saturated graphs, 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.

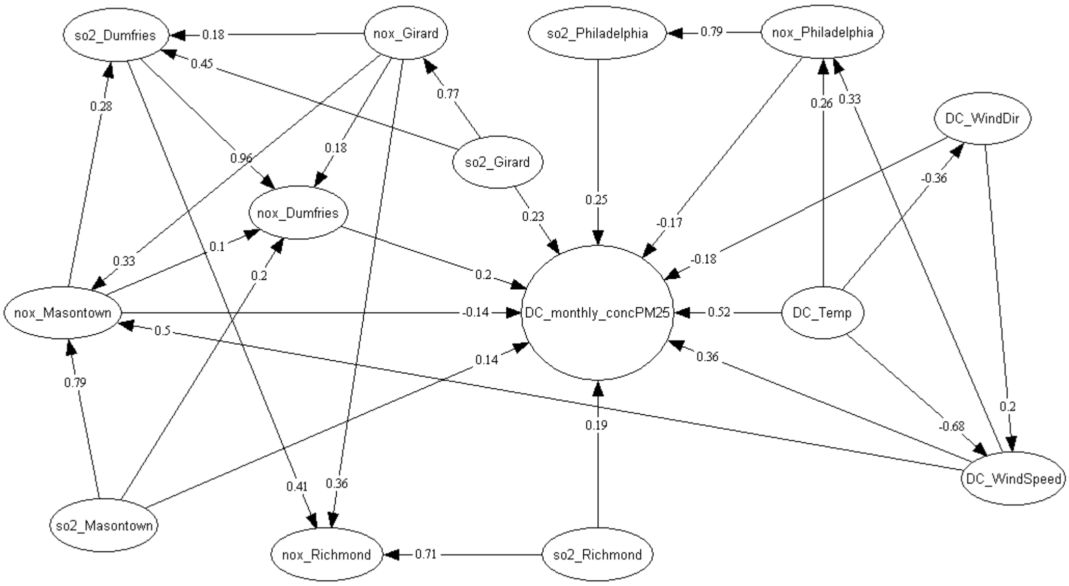


Figure 14: BBN on 14 nodes with 30 arcs.

5 Alternative Ways to Calculate the Correlation Matrix of a BBN

In both learning the structure of the BBN and the conditioning step, which was briefly presented in Section 1, an important operation is calculating the correlation matrix from the partial correlations specified. To do so, we are repeatedly using equation 2.1. When working with 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 2.1. It is known that a BBN induces a (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 $\rho_{XY;Pa(Y)}$.

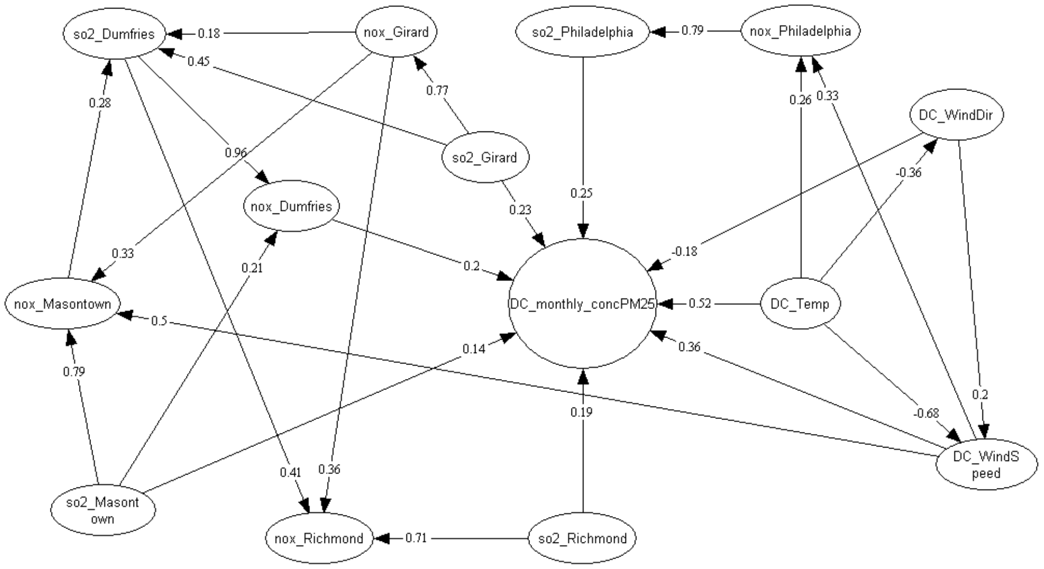


Figure 15: BBN on 14 nodes with 26 arcs.

5.1 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. All definitions presented in this section can be found in the literature, e.g. [27]. Capital letters, e.g. X, denote a single variable. Sets of variables are denoted in bold, e.g. **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.: $An(X) = X \cup \bigcup\limits_{Y \in Pa(X)} An(Y)$. 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 5.1. *If **A**, **B** and **C** are three disjoint subsets of nodes in a BBN, then **C** is said to d-separate **A** from **B** if there is no path between a node in **A** and a node in **B** along which the following two conditions hold:*

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

Figure 16 explains the above definition graphically.
If a path satisfies the conditions above, it is said to be *active*. Otherwise it is said to be blocked by **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 [34] 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 5.1. *$D_{sep}(X;Y|\emptyset)$ implies the absence of a colliderless path between X and Y.*

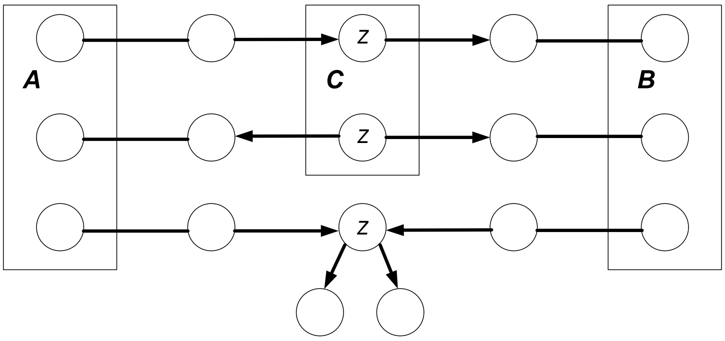


Figure 16: D-separation of **A** & **B** by **C**.

5.2 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, [27] 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 5.1. *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 5.1 it follows that $An(X) \cap An(Y)$ d-separates X from Y, hence $X \perp Y|An(X) \cap An(Y)$. □

However the intersection of the ancestors of X and Y may contain more variables than $Pa(Y)$. In this case it is more convenient to calculate ρ_{XY} starting from $\rho_{XY;Pa(Y)} = 0$.

Proposition 5.2. *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 5.1 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 5.3. *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:

$$P(Y|\mathbf{A}, \mathbf{B}) = \frac{P(Y, \mathbf{A}, \mathbf{B})}{P(\mathbf{A}, \mathbf{B})} = \frac{P(Y, \mathbf{A}|\mathbf{B})\, P(\mathbf{B})}{P(\mathbf{A})\, P(\mathbf{B})} = \frac{P(Y, \mathbf{A})}{P(\mathbf{A})} = P(Y|\mathbf{A}). \tag{5.1}$$

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 [6; 35]). 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 5.1 it is enough to consider the boundary of this intersection.

Proposition 5.4. *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:*

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

Proof. By symmetry, it suffice to prove only one of the relations above. Let $\mathbf{C} = An(X) \cap An(Y)$ and $\mathbf{C}^* = bd_X(An(X) \cap An(Y))$. 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 $\mathbf{C}^*$ is a separator for X and Y. □

Corollary 5.1. *In the context of the previous proposition:*

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

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

If we are in the conditions of Proposition 5.3 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 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
- $14 \perp 12 | bd_{12} \left(An(12) \cap An(14) \right)$.

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

It is clear that with the above results we can reduce the use of equation 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.

6 Conclusions and Future Research

In this paper we have described a method for mining ordinal multivariate data using non-parametric BBNs. The main advantage of this method is that it can handle a large number of continuous variables, without making any assumptions about their marginal distributions, in a very fast and efficient way. Inferring the structure of a BBN from data requires a suitable measure of multivariate dependence. The discussion in this paper led 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 independent, and attains a minimum value of 0 if there is linear dependence between the variables. As mentioned and motivated previously, we actually work with the determinant of the rank correlation matrices. The determinant of the rank correlation matrix is not such an intuitive measure. 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. Another open issue related to this topic is to perform more reliable statistical tests for the two validation steps of our approach.

References

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

[2] J. Cheng, D. A. Bell, W. Liu, An algorithm for bayesian network construction from data, Artificial Intelligence and Statistics.

[3] D. Chickering, D. Geiger, D. Heckerman, Learning bayesian networks is np-hard, Technical Report MSR-TR-94-17, Microsoft Research.

[4] R. Cooke, Markov and entropy properties of tree and vine-dependent variables, in: Proceedings of the Section on Bayesian Statistical Science, American Statistical Association, 1997.

[5] G. Cooper, E. Herskovits, A bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992) 309–347.

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

[7] N. Friedman, M. Goldszmidt, Discretizing continuous attributes while learning bayesian networks., in: In Proc. ICML, 1996.

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

[9] A. M. Hanea, D. Kurowicka, Mixed non-parametric continuous and discrete bayesian belief nets, Advances in Mathematical Modeling for Reliability ISBN 978-1-58603-865-6 (IOS Press).

[10] D. Heckerman, A tutorial on learning bayesian networks, Technical Report MSR-TR-95-06, Microsoft Research.

[11] D. Heckerman, D. Geiger, Learning bayesian networks: a unification for discrete and gaussian domains., UAI (1995) 274284.

[12] D. Heckerman, D. Geiger, D. Chickering, Learning bayesian networks: the combination of knowledge and statistical data, Machine Learning Journal 20(3).

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

[14] H. Joe, Multivariate concordance, Journal of Multivariate Analysis 35 (1990) 12–30.

[15] G. John, P. Langley, Estimating continuous distributions in bayesian classifiers., UAI (1995) 338345.

[16] M. Kendall, A. Stuart, The advanced theory of statistics, Charles Griffin & Company Limited, London, 1961.

[17] D. Kurowicka, Techniques in Representing High Dimensional Distributions, PhD Dissertation, Delft Institute of Applied Mathematics, 2001.

[18] D. Kurowicka, R. Cooke, Uncertainty Analysis with High Dimensional Dependence Modelling, Wiley, 2006.

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

[20] W. Lam, F. Bacchus, Learning bayesian belief networks: an approach based on the mdl principle, Computational Intelligence 10 (1994) 269–293.

[21] D. Lawley, M. Maxwell, Factor Analysis as a Statistical Method, Butterworths Mathematical Texts, London, 1963.

[22] D. Margaritis, Distribution-free learning of bayesian network structure in continuous domains, Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, 2005.

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

[24] O. Morales-Napoles, D. Kurowicka, R. Cooke, D. Ababei, Continuous-discrete distribution free bayesian belief nets in aviation safety with UniNet, Technical Report TU Delft.

[25] R. Morgenstern, W. Harrington, J. Shis, R. Cooke, A. Krupnick, M. Bell, Accountability analysis of title IV of the 1990 clean air act amendments. An approach using bayesian belief nets., in: Annual Conference, Health Effects Institute, Philadelphia, 2008.

[26] R. Nelsen, An Introduction to Copulas, Lecture Notes in Statistics, Springer- Verlag, New York, 1999.

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

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

[29] K. Pearson, Mathematical contributions to the theory of evolution, Biometric Series. VI.Series.

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

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

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

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

[34] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, Springer-Verlag, New York, 1993.

[35] J. Whittaker, Graphical Models in applied multivariate statistics, John Wiley and Sons, Chichester, 1990.

[36] G. Yule, M. Kendall, An introduction to the theory of statistics, Charles Griffin & Co. 14th edition, Belmont, California, 1965.