# Official Statistics Data Integration Using Copulas

Luciana Dalla Valle*

School of Computing and Mathematics, Plymouth University

## Abstract

The aim of this paper is to propose a novel approach to integrate financial information, incorporating the dependence structure among the variables. The approach is based on two types of graphical models: vines and non parametric Bayesian belief nets (NPBBNs). Vines are undirected graphs, representing pair copula constructions, which are used to model the dependence structure of a set of variables. NPBBNs are directed graphs, that use pair copulas to model the dependencies, and allow for diagnosis and prediction via conditionalization. This approach allows to aggregate information and to calibrate the results obtained with different sources of data. The illustrated methodologies are applied to two financial datasets, the first one containing data collected through a survey and the second one containing official statistics data.

***Keywords***: *Data Integration, Official Statistics, Copulas, Vines, non parametric Bayesian belief nets.*

*email: luciana.dallavalle@plymouth.ac.uk

# 1 Introduction

The problem of data integration has become an important issue in the latest years, due to the growth of the number of available data sources and to the increase in data quality standards.

In particular, one of the challenges faced by statistical scientists is to balance results obtained from specific datasets with those obtained from official statistics. When these results differ substantially, it is difficult to interpret and integrate them. Using a method to aggregate these results is fundamental in order to obtain reliable analyses. As illustrated by Foresti et al. [12], the matching of public with private databases is crucial for implementing new analyses that are functional to a new approach to business.

Recently the advances in technology and communications have increased the availability of sources of information and large databases. For this reason multivariate modeling is of fundamental interest and new methods to manipulate high quantities of data have become essential. Unfortunately, high-dimensional modeling with data characterized by complex dependence patterns can be quite challenging.

Copulas have proven to be very promising in statistical problems where data include a number of high-dependent variables. Vine copulas, in particular, are extremely flexible in high-dimensional cases, allowing the specification of various types of non-linear dependencies. However, vines interpretation can be hard when used for prediction, due to the complexity of its undirected structure. Moreover, the use of vines becomes computationally cumbersome when the dimension of the dataset exceeds the order of tens of variables. Non parametric Bayesian belief nets (NPBBNs) have demonstrated adequate for dimensionality of hundreds of variables, with fast computational times. NPBBNs require no assumption on the distributions of the marginals, unlike parametric Bayesian belief nets, and the relationships among variables are specified through copulas. NPBBNs allow a straightforward interpretation of the casualties, thanks to their directed structure. Therefore, NPBBNs can be easily used for prediction via conditionalization, since they clearly show the flow of influence among variables.

The aim of this paper is to use vines to identify the dependence structures of multidimensional datasets, where the variables exhibit a high degree of correlation. Moreover, we will present an innovative approach to integrate and calibrate the information of different datasets, conditionalizing NPBBNs for predictive and diagnostic reasoning. This methodology is applied to two financial datasets. The first dataset includes information of a sample of Italian firms, collected through a survey, while the second dataset contains publicity available data of the Italian national stock exchange (FTSE-MIB). Both datasets include the sales variable, which is the main focus of our analysis.

The remainder of this paper is organized as follows: in Section 2 we present an overview of the existing literature; Section 3 illustrates the characteristics of the proposed novel approach, introducing the theoretical framework, illustrative examples and simulation studies; finally concluding remarks are given in Section 4.

## 2 Literature overview

The existing literature about data integration ranges from traditional models based on regression and linear dependencies, to more complex models allowing to express non-linear dependencies and causal relationships. In the analysis of data characterized by a complex dependence structure, traditional methods often fail to capture the actual relationships among variables and may produce biased results for prediction. However, simple models like multivariate regression are still quite popular. For example, Foresti et al. [12] used OLS to identify the determinants of sales growth, applying it to several integrated private databases. Copulas, introduced by Sklar in 1959 [28], have become very popular in finance, and have been applied to a wide variety of fields, like biology, medicine, social sciences and sampling theory. They allow to calculate the joint multivariate distribution from the marginals, incorporating their dependence structure. The main advantage of copulas is their flexibility, since the marginals may be described by any type of distribution and the various classes of copulas are able to accommodate several types of dependencies. Therefore, copulas can be successfully used to model datasets with complex and non-linear dependence structures, and to aggregate data from different sources. However, while in the bivariate case copulas can be effectively used for dependence modeling, in the multivariate case (typically, when the number of variables exceeds 3) copulas' flexibility is greatly reduced, since the choice of the families is limited to the elliptical copulas (Normal and Student's t). Recently pair copula constructions and their graphical representation, vines, have been introduced by Aas et al. [1], to overcome the lack of flexibility of copulas in high-dimensional cases. Vines have been applied to a variety of financial problems, like in Czado et al. [9], to model the dependencies of US exchange rates, or in Brechmann and Czado [5], to analyze the Euro Stoxx 50. The main advantage of copulas and vines is therefore the ability to model complex dependence structures of variables in a flexible way, and to use it to integrate different scenarios and results obtained with different data. However, the use of vines is not always easy, since their visualization and interpretation is not straightforward.

Probabilistic graphical models are another powerful statistical tool that aims at modeling the dependencies among variables. These models are used to represent multivariate densities via a combination of a qualitative graph structure that encodes independencies and local quantitative parameters. Penny and Reale [26] used graphical models in official statistics to identify the relevant components in a saturated structural VAR model for the quarterly gross

domestic product, that is the aggregation of a large number of economic time series. More recently, Vicard and Scanu [29] applied Bayesian networks to official statistics, showing that the use of post-stratification allows integration and missing data imputation.
However, Bayesian belief networks (BBNs) are affected by some limitations. The main classes of BBNs are in fact discrete, normal or discrete-normal, where discrete BBNs are limited to small-sized datasets, and normal BBNs are limited by the joint normality assumption. For this reason, researchers proposed alternative methodologies. Elidan [11] for example points out the need for a synergy between the copula framework and the field of machine learning. To this end, Kurowica and Cooke [20] and Hanea et al. [15] introduced continuous non parametric Bayesian belief nets, using copulas to realize rank correlations in directed acyclic graphs. This new approach is based on nonparametric statistical inference and elicited expert knowledge to understand the dependencies among the variables, and uses conditionalization for diagnosis and prediction.
An alternative approach, that will not be analyzed in the present work, is given by pair copula Bayesian networks, introduced by Bauer and Czado [2], who conversely focused their attention to parametric likelihood inference and data-driven structure estimation.

# 3 The proposed approach

## 3.1 Theoretical framework

### 3.1.1 Copulas and Vines

The copula is a function that allows to bind together a set of marginals, considering their dependence structure, to obtain the joint multivariate distribution.
More formally, suppose that $U_1, \ldots, U_d$ are random variables uniformly distributed on $[0,1]$. Then, a $d$-dimensional copula $C : [0,1]^d \to [0,1]$ is a multivariate distribution function defined on the unit cube $[0,1]^d$, with uniformly distributed marginals:

$$C(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d).$$

According to Sklar's theorem [28], any joint multivariate distribution $F(x_1, \ldots, x_d)$ of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ can be represented as a copula of its univariate marginals $F_1(x_1), \ldots, F_d(x_d)$, via the following expression

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right).$$

For continuous marginals $F_1, \ldots, F_d$, the copula $C(\cdot, \ldots, \cdot)$ is unique and it is defined through Nelsen's corollary [24]

$$C(x_1, \ldots, x_d) = F(F_1^{-1}(x_1), \ldots, F_d^{-1}(x_d)).$$

The converse statement is also true, since any copula taking any marginal distribution as its arguments defines a $d$-dimensional cumulative distribution function.
The corresponding copula density is given by

$$c(u_1,\ldots,u_d) = \frac{\partial^d C(u_1,\ldots,u_d)}{\partial u_1 \cdots \partial u_d}.$$

The joint density function of a random vector $\mathbf{X} = (X_1,\ldots,X_d)$ is therefore, using the chain rule,

$$f(x_1,\ldots,x_d) = c(F_1(x_1),\ldots,F_d(x_d)) \cdot f_1(x_1)\cdots f_d(x_d) \tag{1}$$

where $c(F_1(x_1),\ldots,F_d(x_d))$ is the $d$-variate copula density.
For example, in the 5-dimensional case $d = 5$ and the (1) becomes

$$f(x_1,\ldots,x_5) = c(F_1(x_1),\ldots,F_5(x_5)) \cdot f_1(x_1)\cdots f_5(x_5)$$

where $c(\cdot,\ldots,\cdot)$ is the appropriate pair-copula density for the transformed variables $F_1(x_1),\ldots,F_5(x_5)$.
Therefore, the copula not only allows to determine the joint multivariate distribution, but also allows to describe the dependencies among the marginals, that are arbitrary and potentially different distributions.
Several families of copulas are available to capture different types of symmetric and asymmetric dependencies among the marginals. The most popular families are the elliptical (i.e. Gaussian and Student's t) and the archimedean copulas (i.e. Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7 and BB8). For an overview of the main types of copulas and their characteristics see for example Joe [19] or Nelsen [24]. However, while the literature on bivariate copulas (or *pair copulas*) has flourished in the recent years, its extension to the multivariate case is rather limited, due to analytical and computational complexity. For this reason the application of copulas to big multivariate datasets characterized by complex patterns of dependence has been rather scarce so far. In order to fill this gap, Bedford and Cooke ([3], [4]) introduced a flexible class of multivariate copulas using bivariate copulas as building blocks. The decomposition of a multivariate copula into bivariate copulas is called *pair copula construction* (PCC) and it allows to express the multivariate distribution of a random vector as a product of pair copulas. Assuming that $f(x_1,\ldots,x_d)$ is the distribution of a random vector $\mathbf{X} = (X_1,\ldots,X_d)$, then it easily factorizes (uniquely up to re-labeling of the variables) into a product of conditional densities

$$f(x_1,\ldots,x_d) = f_d(x_d) \times f_{d-1|d}(x_{d-1}|x_d) \times \ldots \times f_{1|2\cdots d}(x_1|x_2,\ldots,x_d). \tag{2}$$

Considering a 5-dimensional distribution, equation (2) correspond to

$$f(x_1,\ldots,x_5) = f_5(x_5) \times f_{4|5}(x_4|x_5) \times f_{3|4,5}(x_3|x_4,x_5) \times \ldots \times f_{1|2\cdots 5}(x_1|x_2,\ldots,x_5). \tag{3}$$

Then, using the (1), the conditional densities of (2) can be decomposed into the appropriate pair copula times a conditional marginal density. More precisely, for a generic element $X_\jmath$ of the vector $\mathbf{X}$ we obtain

$$f_{x_\jmath|\mathbf{v}}(x_\jmath|\mathbf{v}) = c_{x_\jmath,v_j|\mathbf{v}_{-j}}(F_{x_\jmath|\mathbf{v}_{-j}}(x_\jmath|\mathbf{v}_{-j}), F_{v_j|\mathbf{v}_{-j}}(v_j|\mathbf{v}_{-j})) \times f_{x_\jmath|\mathbf{v}_{-j}}(x_\jmath|\mathbf{v}_{-j}), \tag{4}$$

where $\mathbf{v}$ is the conditioning vector, $v_j$ is an generic component of $\mathbf{v}$, $\mathbf{v}_{-j}$ is the vector $\mathbf{v}$ without the component $v_j$, $F_{x_\jmath|\mathbf{v}_{-j}}(\cdot|\cdot)$ is the conditional distribution of $x_\jmath$ given $\mathbf{v}_{-j}$, and $c_{x_\jmath,v_j|\mathbf{v}_{-j}}(\cdot,\cdot)$ is the conditional pair copula density. For example, the second factor, $f_{4|5}(x_4|x_5)$, in the right-hand side of (3) can be easily decomposed into the pair-copula $c_{4,5}(F_4(x_4), F_5(x_5))$ and a marginal density $f_4(x_4)$:

$$f_{4|5}(x_4|x_5) = c_{4,5}(F_4(x_4), F_5(x_5)) \cdot f_4(x_4).$$

For the third factor in the right-hand side of (3) one of the possible decompositions, using the (4), is

$$f_{3|4,5}(x_3|x_4, x_5) = c_{3,4|5}(F_{3|5}(x_3|x_5), F_{4|5}(x_4|x_5)) \cdot f_{3|5}(x_3|x_5),$$

for the appropriate pair copula $c_{3,4|5}$, applied to the transformed variables $F_{3|5}(x_3|x_5)$ and $F_{4|5}(x_4|x_5)$.

The $d$-dimensional joint multivariate distribution function can thus be expressed as a product of pair copulas by recursively using equation (4) in equation (2). Since in the (4) the conditional distributions of the form $F_{x|\mathbf{v}}(\cdot|\cdot)$ are not directly observable, they are calculated using Joe's result [18]

$$F_{x|\mathbf{v}}(x|\mathbf{v}) = \frac{\partial C_{x,v_j|\mathbf{v}_{-j}}(F(x|\mathbf{v}_{-j}), F(v_j|\mathbf{v}_{-j}))}{\partial F(v_j|\mathbf{v}_{-j})}. \tag{5}$$

If the conditioning set $\mathbf{v}$ is univariate, $\mathbf{v} = v$ and expression (5) can be written as

$$F(x|v) = \frac{\partial C_{x,v}(x, y, \boldsymbol{\theta})}{\partial v} = h(x, v, \boldsymbol{\theta}), \tag{6}$$

where $\boldsymbol{\theta}$ denotes the set of parameters of the copula (measuring the dependencies among the marginals), and $F(x|v)$ is named the *h function*. The forms of the *h functions* for the main classes of copulas are given in Aas et al. [1] and in Czado et al. [9]. For example, $F_{3|5}(x_3|x_5)$ can be determined using expression (6) as follows

$$F_{3|5}(x_3|x_5) = \frac{\partial C_{3,5}(F_3(x_3), F_5(x_5))}{\partial F_5(x_5)}.$$

Therefore, under the assumption of absolutely continuous distributions, a multivariate density can be expressed as a product of pair copulas acting on several different conditional distributions, obtaining a PCC.

PCCs can be represented through a graphical model called *regular vine* (R-vine). An R-vine $\mathcal{V}(d)$ on $d$ variables is a nested set of trees (connected acyclic graphs) $T_1, \ldots, T_{d-1}$, where the variables are represented by nodes linked by edges, each associated with a certain pair copula in the corresponding PCC. The edges of tree $T_j$ are the nodes of tree $T_{j+1}$, $j = 1, \ldots, d-1$. In an R-vine, if two edges of tree $T_j$ share a common node, they are represented in tree $T_{j+1}$ by nodes joined by an edge. Note that there are many different orderings of the variables yielding different R-vines. We can distinguish two subclasses of regular vines, *canonical* or *C-vines* and *D-vines*, each of them giving a specific way of decomposing the density. A C-vine is an R-vine whose trees are all stars, since each tree $T_j$ has a unique node that is connected to $d - j$ edges. This type of vine is particularly suitable for datasets where a variable is known to be a key variable that should be located at the root of the C-vine. Conversely, a D-vine is an R-vine where all vertices in tree $T_1$ are adjacent to at most two other vertices. Figure 1 shows a C-vine and a D-vine with five variables, in the left and in the right panel, respectively. Both vines consists of four trees $T_j$, $j = 1, \ldots, 4$. Each edge corresponds to a pair copula density (possibly belonging to different families) and the edge label corresponds to the subscript of the pair copula density, e.g. edge 34|12 corresponds to the copula density $c_{34|12}(\cdot)$. For the C-vine represented in Figure 1 the joint density is given by

$$f(x_1, \ldots, x_5) = \prod_{j=1}^{5} f_j(x_j) \times c_{12} \times c_{13} \times c_{14} \times c_{15} \times c_{23|1} \times c_{24|1} \times c_{25|1} \times c_{34|12} \times c_{35|12} \times c_{45|123}.$$

Note that in the previous equation the notation has been simplified, setting $c_{ab} = c_{ab}(F(x_a), F(x_b))$.

`Figure 1 approximately here`

Therefore, any multivariate positive density can be decomposed according to the corresponding PCC and it can be represented through a vine.

Considering the R-vine estimation, the vine structure as well as the copula parameters have to be specified. In order to select a suitable R-vine decomposition, a sequential approach is generally adopted, specifying the first tree and then proceeding similarly for the following trees. For selecting the structure of each tree, we followed the approach suggested by Aas et al. [1] and developed by Dißmann et al. [10], using the maximal spanning tree algorithm. This algorithm defines a tree on all nodes (named spanning tree), which maximizes the sum of absolute pairwise dependencies, measured, for example, by Kendall's $\tau$. This specification allows to capture the strongest dependencies in the first tree and to obtain a more parsimonious model.

Then, given the selected tree structure, a copula family for each pair of variables is selected using the Akaike Information Criterion (AIC), or the Bayesian Information Criterion (BIC). This choice is made amongst a large set of families, comprising elliptical copulas (Gaussian and Student t) as well as archimedean copulas (Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7 and BB8) and their rotated versions, to cover a large range of possible dependence structures.

Note that conditional independence between variables may reduce the number of levels of the pair copula decomposition, and hence simplify the construction (removing edges in the R-vine). Therefore, an independence test (see Genest and Favre [13]) is performed on each pair of variables.

After specifying the vine structure and the copula families, the copula parameters $\boldsymbol{\theta}$ are then estimated using the maximum likelihood method, as illustrated by Aas et al. [1]. Alternatively, Bayesian inference can be adopted to estimate the parameters, as shown in Min and Czado [23].

Finally, Joe's formula (5) is employed to calculate the conditional distributions $F(x|\mathbf{v})$, that are used as input for the next trees.

Note that, as already mentioned, the R-vine estimation procedure is repeated for all the trees, until the R-vine is completely specified.

### 3.1.2 Continuous Non Parametric Bayesian Belief Nets

R-vines represent very flexible tools to model complex patterns of dependence among marginal distributions through a rich class of pair copula families. R-vines have been successfully applied to datasets with dimensionality of at most tens of variables, as illustrated by Brechmann and Czado [5] who investigated the dependence structure of the Euro Stoxx 50 and analyzed a 52-dimensional dataset. However, with datasets of dimensionality of hundreds of variables, R-vines become computationally intractable. For this reason, NPBBNs were introduced in Kurowicka and Cooke [20] and extended in Hanea et al. [15].

Bayesian belief nets (BBNs) are directed acyclic graphs (DAGs) whose nodes represent variables and the arcs represent causal relationships between the variables. These variables are associated to conditional probability functions that, together with the DAG, are able to provide a compact representation of high-dimensional distributions. For an introduction and for more details about the definitions and main results see, for example, Cowell [7], Jensen [16], Jensen [17] or Pearl [25].

The most popular classes of BBNs are discrete, normal or discrete-normal. Recall that the direct predecessors of a node, corresponding to a variable, are called *parents*, while the direct descendants of a node are called *children*. Then, the nodes of discrete BBNs are discrete random variables, that are represented as marginal distributions for parent nodes and conditional probability tables for child nodes. In normal BBNs the nodes form a joint gaussian distribution, where the influence of the parent nodes on a child node is interpreted as partial regression coefficients where the child is regressed on the parents (see Shachter and Kenley [27]). However, discrete BBNs are only suitable to datasets of limited size and complexity, while normal BBNs are limited by the assumption of joint normality. In order to overcome this limitations, Kurowicka and Cooke in [21] introduced NPBBNs, where distributions can conform to any parametric form and the relationships among variables are defined through R-vines. 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 conditional independence statements encoded in the graph allow to write the joint density as

$$f(x_1,\dots,x_d)=\prod_{j=1}^{d} f_{x_j|\boldsymbol{Pa(j)}}(x_j|\boldsymbol{Pa_j})$$

where $f_{x_j|\boldsymbol{Pa(j)}}$, with $j=1,\dots,d$, is the conditional probability function associated to node $j$, that corresponds to variable $X_j$, and $\boldsymbol{Pa_j}$ is the set of all $j$'s parents. Hence, the nodes are associated with continuous invertible distributions, while each arc is represented by a (conditional) rank correlation. More precisely, for each node $j$ with parents $j_1,\dots,j_{p(j)}$, the arc $j_{p(j)-k}\longrightarrow j$ is associated with the (conditional) rank correlation

$$\begin{cases} r_{j,j_{p(j)}} & \text{if} \quad k=0 \\ r_{j,j_{p(j)-k}|j_{p(j)},\dots,j_{p(j)-k+1}} & \text{if} \quad 1\le k\le p(j)-1 \end{cases}$$

where the assignment is vacuous if the considered node does not have any parents, that is $\{j_1\cdots j_{p(j)}\}=\varnothing$. Therefore, every arc in the NPBBN is associated with a (conditional) rank correlation between parent and child, according to a protocol presented, for example, in Kurowicka and Cooke [20].

Figure 2 approximately here

As an example, we illustrate the assignments for the DAG of the NPBBN depicted in Figure 2, as explained in Hanea [14]. The first step is to construct a sampling order of the nodes and to index the nodes according to it. Between the two possible orderings (1, 2, 3, 4) or (1, 3, 2, 4), we choose the first one. The second step is the factorization of the joint following the sampling order, that in our case gives us

$$P(1)P(2|1)P(3|21)P(4|321).$$

The following step consists in highlighting the nodes in each conditioning set that are not parents of the conditioned variable, obtaining

$$P(1)P(2|1)P(3|1\underline{2})P(4|32\underline{1}), \tag{7}$$

where the underscored variables are not necessary in sampling the conditioned variable and can be omitted. Therefore, the rank correlations to be assigned to the arcs in Figure 2 are $\{r_{12}, r_{13}, r_{43}, r_{42|3}\}$. These assignments uniquely determine the joint distribution and are algebraically independent, as proved in Kurowicka and Cooke [20]. For each term of the factorization (7) a D-vine is built, whose (conditional) rank correlations exactly correspond to those of the NPBBN (see Hanea et al. [15]). The (conditional) rank correlations and the marginal distributions needed to completely specify the NPBBN can be retrieved from data or elicited from experts.

In order to sample the joint distribution, the same procedure used for D-vines can be employed (for more details about sampling a D-vine, see Aas et al. [1]). However, as explained in Hanea [14], the order of the variables in successive D-vines might not be same. Sampling involves the calculation of complex conditional cumulative distribution functions, that may be not analytically tractable. Since their numerical evaluation for every sample is computationally heavy, it is recommended to use the normal copula to realize the conditional rank correlations. Restricting the choice on normal copulas allows us to avoid the computational problems associated to NPBBNs, due to the nice properties of this particular type of copula (for more details see Kurowicka and Cooke [20]).

Continuous NPBBNs are in many ways similar to vines. They both express the dependence structure among the marginal distributions through conditional copulas, associated to the arcs of the corresponding graph. However, while NPBBNs are described by directed graphs, in vines the arcs are undirected, and different conditional independence statements are specified. More specifically, as explained by Hanea [14], in NPBBNs conditional independence is expressed by the absence of an arc connecting two nodes, but the presence of an arc does not guarantee dependence. On the contrary, in R-vines fully connected graphs represent conditional dependence statements, but conditional independence does not always correspond to a missing arc.

## 3.2 Illustrative examples

### 3.2.1 Analysis of Assolombarda data

After introducing the methodology, we now apply vines and NPBBNs to a dataset of Italian firms. The information do not come from official sources, but were collected by an association called Assolombarda. This is an Italian association of about 5,000 firms located in the province of Milan and in other provinces of the north of Italy, and represents manufacturing and service companies. The associated firms employ about 300,000 workers locally and several hundred thousands in the whole country. Assolombarda periodically collects data through questionnaires sent to the associated firms, in order to gather information about the economic climate, firms' activity and production, and the number and types of employees. The data analyzed here contain information collected through one of the association surveys in 2007, and it is about 167 firms located in the provinces of Milan and Lodi. The variables in the dataset are

- *sales*: firm annual turnover;
- *emp*: average number of employees;
- *rise*: number of managers receiving wage rise;
- *rise2*: number of managers that will receive wage rise in the following year;
- *prom*: number of employees gaining a promotion;
- *horiz*: number of employees involved in horizontal movements;
- *ext*: number of people employed in the external market;
- *grad*: number of newly-graduated employees;
- *qual*: number of newly-qualified employees.

Therefore, the dimensionality of the dataset is $d = 9$, which makes the traditional multivariate copula approach unfeasible, while vines and NPBBNs are perfectly adequate.

**Canonical Vine**

Since in the dataset "sales" is the target variable and dominates the dependencies with all the remaining variables, we decided to use C-vines amongst other vine structures, and to set "sales" as the root node.

After choosing the C-vine structure, we fitted the vine copula model to the data. Firstly, we used the inverse transformation method to transform the original data into pseudo-observations, lying in the interval $[0, 1]$. In particular, we employed the *canonical maximum likelihood* (CML) method, that, using the empirical cumulative distribution functions of each marginal to calculate the pseudo-observations, does not require the specification of the forms of the marginal distributions.

In order to select an appropriate C-vine copula model, we followed the method described in section 3.1.1. For each variable pair we performed the bivariate asymptotic independence test (Genest and Favre [13]), where the rejection of the null hypothesis denotes a strong dependence between the selected marginals. Subsequently, for the variable pairs that were not identified as independent by the Genest and Favre test, we selected the appropriate pair copula families. More precisely, for given bivariate copula data, we chose the copula among Elliptical, Archimedean and rotated Archimedean families using the AIC, although the BIC produced very similar results. A preference towards the AIC over BIC in vines has been expressed by other authors in the literature, like for example Brechmann and Czado [5]. Finally, we estimated the parameters of each copula with the maximum likelihood method (see Aas et al. [1]).

`Table 1 approximately here`

The parameter estimates of the C-vine for the Assolombarda dataset are listed in Table 1. The Table displays the list of pair copulas in the trees of the C-vine, the selected copula family and the copula parameters (that can be one or two according to the type of copula). From the copula families selected, we see evidence of different types of asymmetric dependence, easily captured by the flexible vine copula approach. Note that only the first and second tree do not have any conditional independent variable pair. In higher order trees the Genest and Favre test selected a number of independent copulas, so that the C-vine structure is simplified and all the dependencies (edges) in the last trees can be removed. In Table 1 e.g. the conditional copula $c_{qual,rise|sales,ext}$ is independent, meaning that the number of newly-qualified employees is not dependent on the number of managers receiving wage rise, given the values of the sales and the number of people employed in the external market.

Observing the strongest dependencies in the unconditional pair copulas in the first tree, we note a high dependence between the sales and the number of employees ($c_{sales,emp}$), and also between the sales and number of managers receiving wage rise ($c_{sales,rise}$). Looking at the conditional copulas instead, we notice a strong relationship between the number of people employed in the external market and in horizontal movements, given the sales value ($c_{ext,horiz|sales}$). The remaining dependence results can be interpreted in a similar way.

Figure 3, obtained with the R package CDVine described in Brechmann and Schepsmeier [6], shows the first two trees of the C-vine tree plot for the Assolombarda data. The squares represent the nodes (variables), while the lines represent the arcs (dependencies). The names of the nodes are written in the squares, and the pair copula families and Kendall's $\tau$ values corresponding to copula parameters are written on the edges. The thicker the grey line the higher the dependence between the variables represented by the nodes.

`Figure 3 approximately here`

In the first tree of the C-vine, depicted in the top panel of Figure 3, the highest dependencies (thickest arcs) are between the sales and the number of employees, and between the sales and number of managers receiving wage rise. The first dependence corresponds to the unconditional pair copula $c_{sales,emp}$, while the second dependence corresponds to $c_{sales,rise}$, as already highlighted in Table 1. In the second tree, represented in the bottom panel of Figure 3, we notice e.g. a strong dependence (thick arc) between the central node "sales, ext" and the node "sales, horiz". This arc corresponds to the pair copula $c_{ext,horiz|sales}$ and it can be interpreted as the relationship between the number of people employed in the external market and in horizontal movements, given the sales value, as mentioned above.

## Non Parametric Bayesian Belief Nets

As we discussed in the previous section, the vine copula model calculated for the Assolombarda data is very useful to understand the relationship among the variables, especially when their dependence pattern is very complex. However, if a specific value of a variable is observed, one could be interested in calculating the new conditional distributions of the remaining variables, in order to understand the effect of one variable on the others. These conditional distributions can be obtained by simulation, but it is not so straightforward to compare them with the unconditional distributions. The main advantage of NPBBNs is their use for inference purposes. The impact of conditionalized variables on the other variables can be easily visualized and compared with the original distributions. The interpretation of the results is also very intuitive, due to the directionality of the arcs. In this context, NPBBNs can be used for prediction, when the effect on a target variable of the remaining variables is observed, or for diagnosis, when the effect of a fixed value of a target variable is observed on the remaining variables.

`Figure 4 approximately here`

Figure 4 shows the NPBBN for the Assolombarda data, that was created with the software UniNet[1]. In the top panel the variables are represented by nodes, while in the bottom panel the variables are represented by histograms, together with their means and standard deviations. As you can notice from the histograms, all the variables are highly right-skewed, making the traditional normality assumption completely unrealistic. Figure 4 illustrates by arcs directed towards the "sales" node the influence on the target variable of the other variables in the dataset. The numbers written on the arrows, denoting the rank correlations, suggest e.g. a high dependence between the number of employees and the sales. Moreover, the oriented arcs in Figure 4 show the relationships among all the other variables, like the influence of both "grad" and "qual" on "ext", meaning that the employees' level of education determines their employment in the external market.

`Figure 5 approximately here`

Figure 5 illustrates two examples of predictive and diagnostic reasoning, respectively in the top and bottom panel. In particular, suppose that we wish to predict the sales value when the number of employees is very high, conditionalizing on the "emp" variable. Thus, we set "emp" to be equal to 32,000, which is much higher than the mean value of this variable (that is 364). In this case the sales value on average goes from 188,080 to 4,745,600, due to the strong dependence between "emp" and "sales". Since the "emp" variable positively influences not only "sales", but also "prom", "rise" and "horiz", the value of all the aforementioned variables rise substantially (top panel of Figure 5). Since "rise" has the strongest conditional rank correlation with "emp", this is the most affected variable by the change in the number of employees, rising from 4.36 to about 36.

Now suppose that we wish to examine the ideal situation in which the "sales" are very high and we want to know which values should the other variables take to obtain this result. Conditionalizing on a high value of "sales", like 700,000, we can use the NPBBNs for diagnosis purposes. As we see in Figure 5, all the variables should rise with "sales", especially "emp" and "ext". Therefore, in order to reach this high sales value, the average number of employees should considerably rise from 364 to 1,076.5 and the average number of people employed in the external market should go from about 5 to 9. For variables with a limited influence on the sales instead, like "horiz", "qual" and "rise2", we note a small increase in their mean figures.

[1] The software is available on http://dutiosc.twi.tudelft.nl/ $\sim$ risk/.

### 3.2.2 Analysis of the FTSE-MIB data

The second application analyses the FTSE-MIB (formerly MIB30) data, an official source. The FTSE-MIB is the benchmark stock market index for the Italian national stock exchange and consists of the 40 most-traded stock classes on the exchange. The dataset analyzed here contains information from the balance sheets of the 40 largest Italian firms belonging to the Italian stock market. For comparison purposes we excluded banks and insurance groups from the original dataset. The data, referring to 2007, are available on the World Wide Web. Specifically, the dimension of the dataset is $d = 8$, and the variables used in our analysis are:

- *sales*: firm annual turnover;
- *emp*: average number of employees;
- *goodwill*: difference between the balance sheet assets and the sum of its intangible assets and equipment at market value;
- *ncas*: non-current financial assets;
- *stocks* : stocks and work in progress;
- *prov*: provisions for liabilities and non-recurring expenses;
- *ncliab*: non-current liabilities;
- *cliab*: current liabilities.

**Canonical Vine**

As in the first example, “sales” is the variable dominating the dependencies of the whole FTSE-MIB dataset. Thus, we chose a C-vine structure to represent the model, with “sales” playing the role of root node. Again, we employed the CML method to obtain pseudo-observations, using empirical cumulative distribution functions for the marginals and avoiding the specification of parametric distributions. Moreover, the Genest and Favre bivariate asymptotic independence test was employed on each pair copula, to check for independence between the corresponding marginals. Then, for each variable pair an appropriate copula family was selected via AIC and, finally, the copula parameters were estimated using the maximum likelihood method.
The parameters estimates of the C-vine for the FTSE-MIB dataset are listed in Table 2.

`Table 2 approximately here`

Table 2 displays the list of pair copulas in the trees of the C-vine, the selected copula family and the copula parameters. In this case only the first tree does not have any conditional independent variable pair. In the second tree $c_{ncliab,ncas|sales}$ has been identified as an independent copula, suggesting the absence of dependence between non current liabilities and assets, conditionally on the value of sales. This is depicted in the second tree of Figure 6 as a separate node, not linked to the remaining nodes.

`Figure 6 approximately here`

Examining the unconditional pair copulas we note a strong dependence between the sales and the number of employees, like in the Assolombarda dataset. This is evident looking at the high value of the dependence parameter in Table 2 for $c_{sales,emp}$, and observing the thick edge linking "sales" and "emp" in the first tree of Figure 6. However, we note that "sales" is quite strongly unconditionally dependent on all the remaining variables, as denoted by high parameter values in Table 2, and as represented by thick edges linking all the nodes in the first tree of Figure 6. On the contrary, the conditional dependencies are rather weak, as demonstrated by small parameter values in Table 2 for the second to the last tree, and thin edges in the second and in the remaining trees (Figure 6). This means that the main dependencies are captured by the unconditional copulas. Therefore, we can say that non-current liabilities are quite strongly correlated to sales ($c_{sales,ncliab}$); however, at the same time non-current liabilities are weakly correlated to provisions for liabilities, given the sales ($c_{ncliab,prov|sales}$).

### Non Parametric Bayesian Belief Nets

Figure 7 shows the NPBBN for the FTSE-MIB data. The top panel depicts the NPBBN with nodes representing the variables, while in the bottom panel the variables are represented by histograms, with their means and standard deviations. Similarly to the Assolombarda data, all the variables are highly right-skewed, as it is clear from the histogram. Here sales are mostly influenced by the values of stocks and non-current labilities, and fairly by the number of employees, while "goodwill" does not seem to have a direct high impact on "sales" (as you may see from the conditional rank correlations on the oriented arcs). Moreover, considering the relationships among the non-target variables, we note that e.g. non-current liabilities not only influence the "sales", but also "goodwill", "ncas" and "stocks".

`Figure 7 approximately here`

This scenario is also summarized in the first four columns of Table 3, that lists the medians, means and standard deviations of the FTSE-MIB variables, as calculated from the original dataset.

`Table 3 approximately here`

Figure 8, top panel, illustrates a scenario where “emp”, “ncas” and “stocks” are set at a low value. More precisely, the number of employees is set at 365, the stock value is 240 and the non-current asset value is 50. As shown in the fifth to seventh columns of Table 3, the value of “sales” decreases from 12,595,630 to 194,510, such that it becomes similar to that of Assolombarda data. Therefore, conditionalizing a NPBBN based on official data, it is possible to set the variables to reflect the status of a particular company and check if the target variable (“sales”) corresponds to the official results. In the bottom panel of Figure 8 we conditionalized the “sales” value to be similar to its mean value in the Assolombarda data (188,000). This scenario is summarized in the last three columns of Table 3, where the new medians, means and standard deviations of all the variables are displayed. Figure 8 and Table 3 suggest what characteristics the firms should have in order to reach a particular level of sales, e.g. a number of employees of more than 2,700 and a stock value of almost 70,000. This approach is very useful to check if a particular company performs similarly to other companies in official studies and to calibrate the results obtained with a specific dataset.

`Figure 8 approximately here`

## 3.3 Simulation study

After defining the structure and estimating the parameters of C-vines and NPBBNs, our aim is now to check the fit of the proposed models to the two illustrated datasets. Since the main purpose of vines and NPBBNs is to model the dependence structure of the marginals and to use it for inference purposes, we focus on how well the considered models replicate the original data, embedding the actual relationships among the variables. Therefore, we generated 1000 simulations of the two datasets using the two models (C-vines and NPBBNs), and we compared the distribution of the original variables with the simulated variables. Since the simulations are generated from the two models with parameters estimated as specified in Section 3.1, they reflect the dependence structure of the marginals. We considered the multivariate t copula as a benchmark, since a multidimensional elliptical copula is the traditional approach to model high-dimensional dependencies, and the t copula in particular is the standard choice for financial data. In order to assess the similarity of the simulated data to the observed data, we performed the Kolmogorov-Smirnov test for the equality of distributions for each simulation. Then, we calculated the median, mean and standard deviation of the p-values of the tests for each variable of the two datasets, as shown in Tables 4 and 5. The closer the median and mean to 1 the better the fit.

`Table 4 approximately here`

`Table 5 approximately here`

Table 4 illustrates the test results for the Assolombarda data, while Table 5 refers to the FTSE-MIB dataset results. Both Tables list the node names in the first column, the p-values median, mean and standard deviations for the multivariate t copula in columns 2 to 4, analogous results for the C-vine in columns 5 to 7, and analogous results for the NPBBN in columns 8 to 10.

As we note from Tables 4 and 5, the C-vine and the NPBBN perform almost always better than the traditional multivariate t copula. In particular, focusing on the sales variable, for the Assolombarda dataset the NPBBN provides the best fit, with an average p-value of 0.83940. For the FTSE-MIB dataset, instead, the C-vine provides the best fit, with an average p-value of 0.87035.

Nevertheless, we need to point out that the main limitation of vines and NPBBNs compared to the standard multivariate elliptical copulas is that they may have a very complex structure, with many parameters to be estimated. In particular, if dependencies in NPBBNs are not expressed by bivariate normal copulas, the computational complexity of the whole structure may become cumbersome. However, the great flexibility of vines and NPBBNs and their graphical representation allows them to effectively describe highly dependent data and to easily interpret the relationship among the variables.

# 4 Conclusion

In this paper we presented a new approach to integrate the information provided by official sources with information provided by other sources.

We used PCCs and their corresponding graphical models, vines, to model the dependance structure of the variables and to calculate the conditional rank correlations. Then, we used NPBBNs to understand the influence of some variables on others and to make predictions and diagnostics.

We firstly applied the illustrated methodologies to a dataset of specific firms, containing information collected by a private association. Then, we applied the methods to a dataset containing official information about the balance sheets of firms in the FTSE-MIB.

The use of vines allowed us to calculate the dependencies among the variables and, in particular, the conditional rank correlations. These correlations were then used to define the arcs of the NPBBNs that were built for the two datasets. In a simulation study NPBBNs performed better than the traditional multivariate t copula, and generally better than the C-vine. The vine performance could be improved by employing the R-vine, instead of the C-vine, since its less constrained structure better adapts to different types of data and various dependencies.

Conditionalizing on a target variable, it is possible to calibrate the values of a specific dataset, considering the dependence structure of the variables. In our case, we can easily calibrate the two datasets via conditionalization and see what characteristics a set of firms should have in order to perform similarly to the firms described in the official data source. In this way we are able to integrate different sources of information.

This study could be extended by including expert opinions and eliciting prior distributions. This prior information could be used to construct the conditional rank correlations in the NPBBNs and could be subsequently updated using observed data.

Furthermore, this approach can be applied to different datasets, as for example those containing categorical variables. This will definitely enhance the applicability of the methods to all types of survey data, and will help to calibrate them with official data. Both vines and NPBBNs allow the specification of discrete variables, however discrete vines have been introduced only recently in the literature and require future investigation and research.

# Acknowledgements

I am grateful to Prof. Ron Kenett for suggesting the topic of this paper and for helpful comments.

# References

[1] Aas, K., Czado, C., Frigessi, A. , Bakken, H. (2009). Pair-copula constructions of multiple dependence, Insurance: Mathematics and Economics, 44, 182–198.

[2] Bauer, A. & Czado C. (2013). *Pair-copula Bayesian networks.* arXiv:1211.5620 [stat.ME].

[3] Bedford, T. & Cooke, R.M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines, *Annals of Mathematics and Artificial Intelligence*, **32**, 245–268.

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

[5] Brechmann, E.C. & Czado, C. (2013). Risk Management with High-Dimensional Vine Copulas: An Analysis of the Euro Stoxx 50. *Statistics & Risk Modeling*, in press.

[6] Brechmann, E.C. & Schepsmeier, U. (2013). Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine, *Journal of Statistical Software*, **52**, 1–27.

[7] Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Sciences, (Springer).

[8] Czado, C. (2010). Pair-Copula Constructions of Multivariate Copulas, in P. Jaworski (Ed.), *Copula Theory and its Applications, Lecture Notes in Statistics*, **198**, Springer, 93–109.

[9] Czado, C., Schepsmeier, U. & Min, A. (2012). Maximum Likelihood estimation of mixed C-vines with application to exchange rates. *Statistical Modelling*, **12**, 229–255.

[10] Dißmann, J., Brechmann, E., Czado, C. & Kurowicka, D. (2011). Selecting and estimating reguilar vine copulae and application to financial returns, *preprint.*

[11] Elidan, G. (2013). Copulas in Machine Learning. *Copulae in Mathematical and Quantitative Finance.* Proceedings of the Workshop Held in Cracow 2012. Springer, 39–60.

[12] Foresti, G., Guelpa, F. & Trenti, S. (2012). Enterprises in a globalised context and public and private statistical setups. *SIS Scientific Meeting 2012.*

[13] Genest, C. & A. C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, **12**, 347–368.

[14] Hanea, A. (2011). Non-Parametric Bayesian Belief Nets versus Vines, In D. Kurowicka & H. Joe (Eds.), *Dependence Modeling. Vine Copula Handbook*, 281–303. Singapore: World Scientific Publishing.

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

[16] Jensen, F. V. (1996). *An Introduction to Bayesian Networks*, London: Taylor and Francis.

[17] Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*, Springer.

[18] Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *IMS lecture notes*, **28**, 120–141.

[19] Joe, H. (1997). *Multivariate model and dependence concepts*, Monographs on Statistics an Applied Probability, **73**, Chapman & Hall, London.

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

[21] Kurowicka, D. & Cooke, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*, Chichester: John Wiley & Sons.

[22] Kurowicka, D. & Cooke, R. M. (2010). Vines and Continuous Non-parametric Bayesian Belief Nets with Emphasis on Model Learning. In K. Böcker (Ed.): *Rethinking Risk Measurement and Reporting*, Risk Books, 295–329.

[23] Min, A. & Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions, *Journal of Financial Econometrics*, **8**, 511–546.

[24] Nelsen, R. B. (2006). An introduction to copulas, Springer-Verlag, New York.

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

[26] Penny, R.N. & Reale, M. (2004) Using graphical modelling in official statistics. *Quaderni di Statistica*, **6**, 31–48.

[27] Shachter, R. D. & Kenley, C. R. (1989). Gaussian Influence Diagrams, *Management Science*, **35**, 527–50.

[28] Sklar, M. (1959): Fonctions de répartition á *n*dimensions et leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

[29] Vicard, P. & Scanu, M. (2012) Applications of Bayesian Networks in Official Statistics. In: A. Di Ciaccio, M. Coli & J. M. Angulo Ibanez (Ed.) *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Springer, 113–123.

# Figures

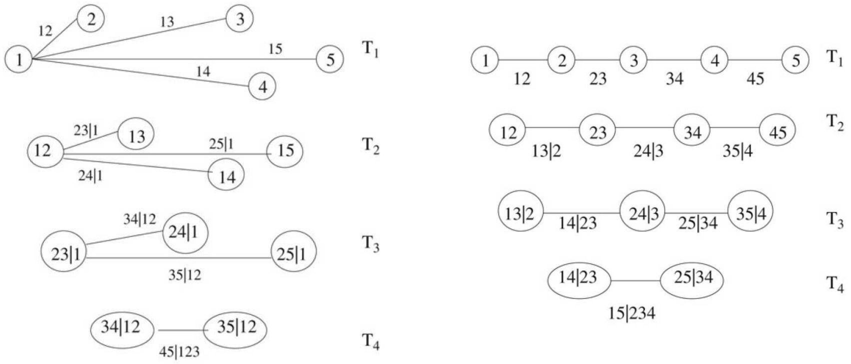


Figure 1: A C-vine (left panel) and a D-vine (right panel) with 5 variables.

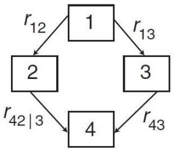
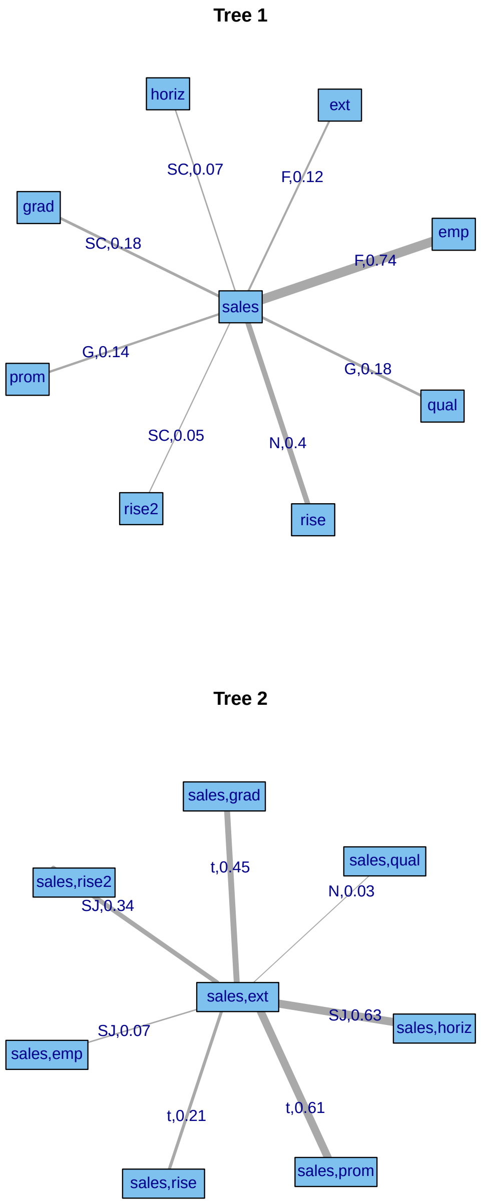


Figure 2: A BBN on four variables with conditional rank correlations assigned to arcs.


Tree 1

Tree 2

Figure 3: First (top) and second (bottom) C-vine trees for the Assolombarda data.

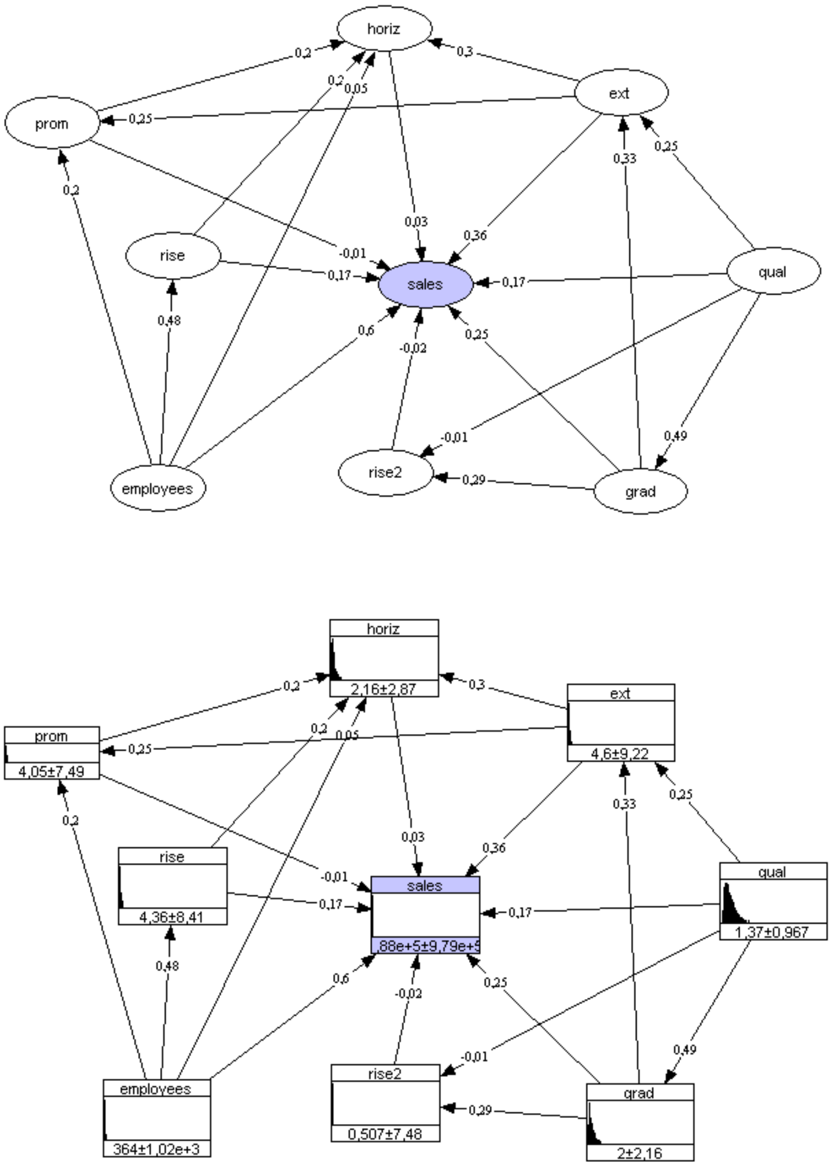


Figure 4: NPBBN for the Assolombarda data. The variables are represented with nodes in the top panel and with histograms in the bottom panel.

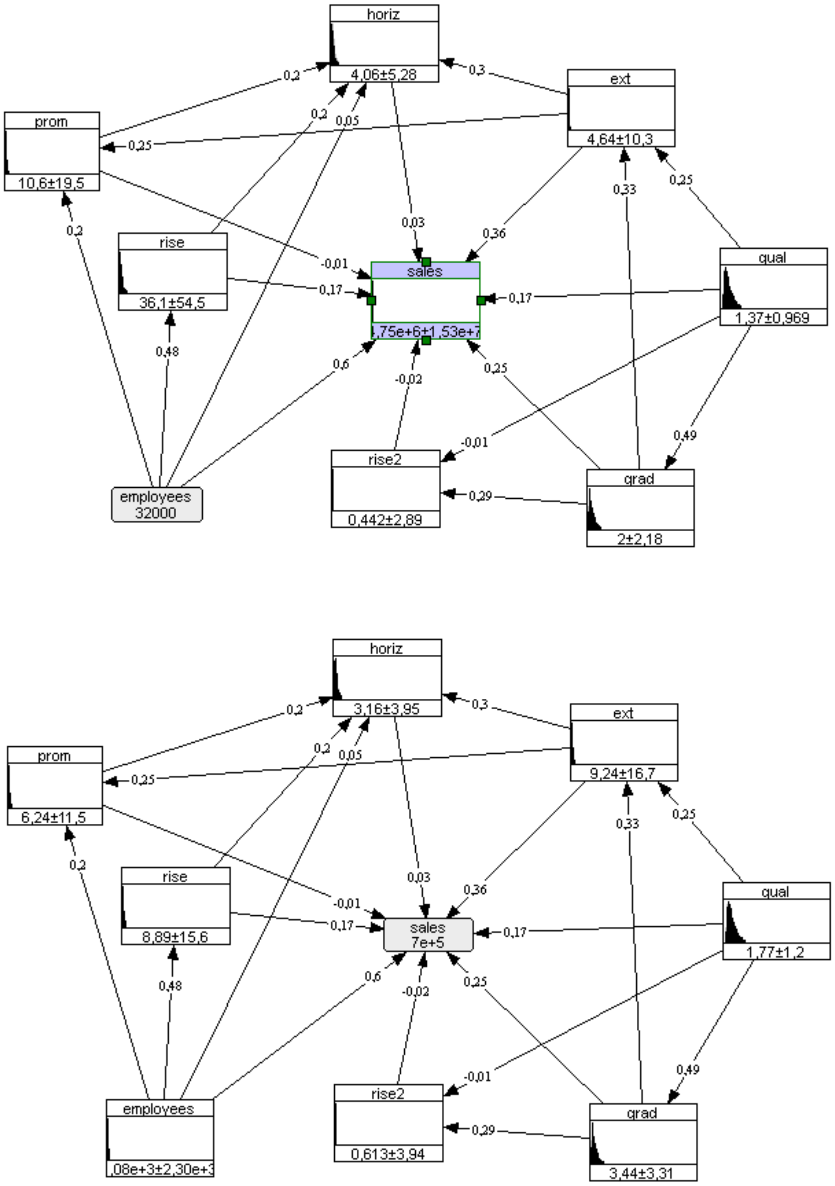


Figure 5: Conditionalized NPBBN for the Assolombarda data. In the top panel the NPBBN is conditionalized for a high value of "emp" (predictive reasoning). In the bottom panel the NPBBN is conditionalized for a high value of "sales" (diagnostic reasoning).

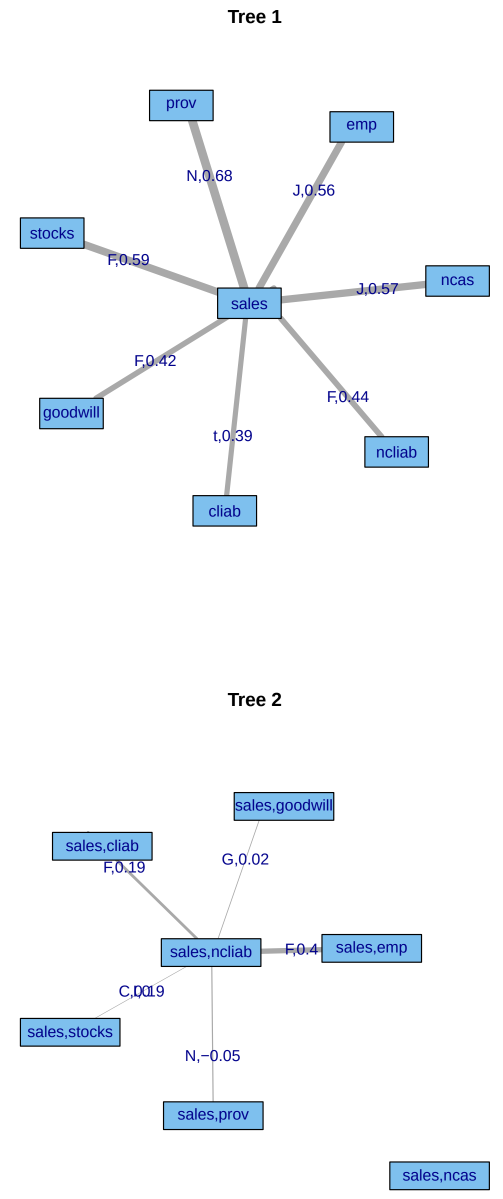


Figure 6: First (top) and second (bottom) C-vine trees for the FTSE-MIB data.

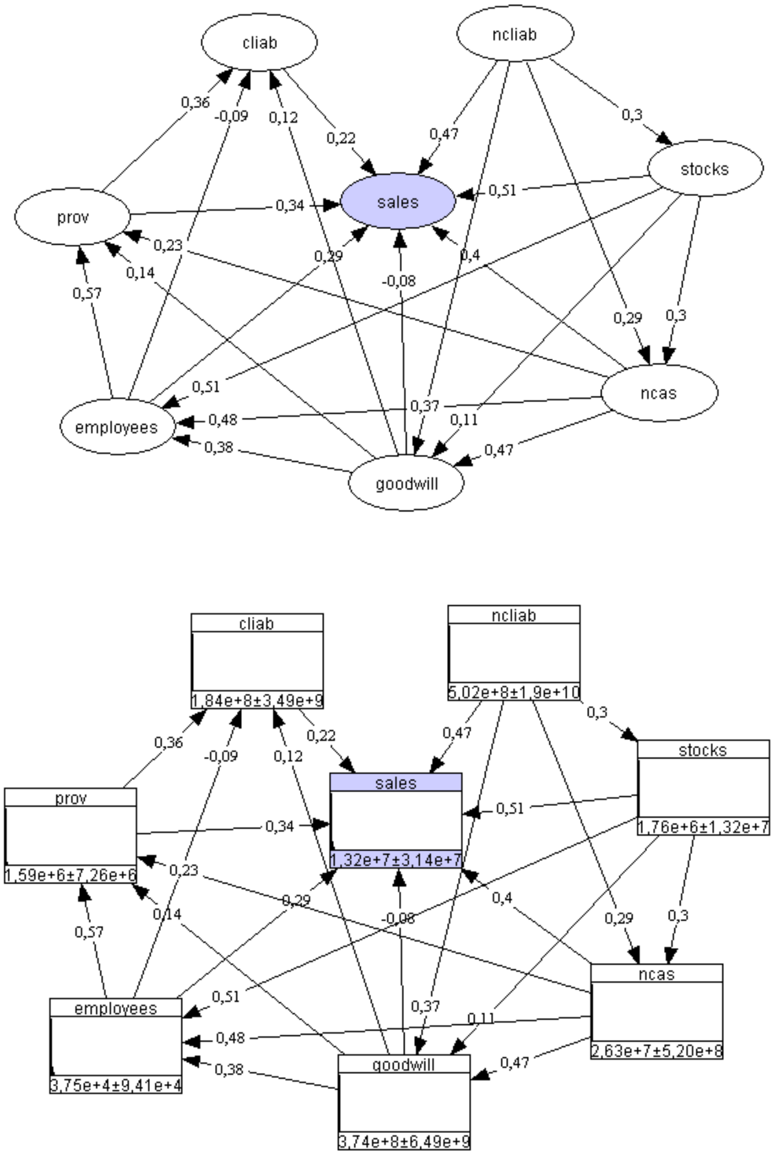


Figure 7: NPBBN for the FTSE-MIB data. The variables are represented with nodes in the top panel and with histograms in the bottom panel.

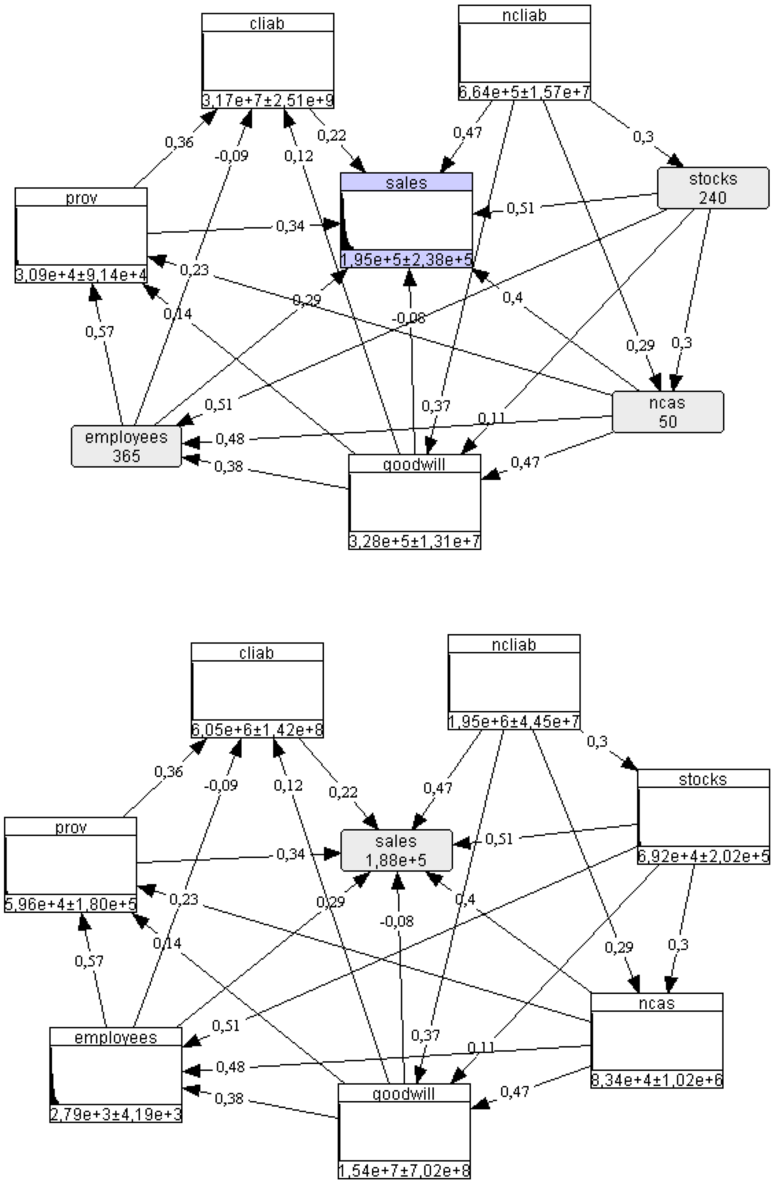


Figure 8: Conditionalized NPBBN for the FTSE-MIB data. In the top panel the NPBBN is conditionalized for a low value of "emp", "ncas" and "stocks" (predictive reasoning). In the bottom panel the NPBBN is conditionalized for a low value of "sales" (diagnostic reasoning).

# Tables

| **Assolombarda data: Parameters of the C-Vine** | | | | |
|---|---|---|---|---|
| *Tree* | *Copulas* | *Family* | *Parameter 1* | *Parameter 2* |
| 1 | $c_{sales,ext}$ | Frank | 1.1225 | 0 |
| | $c_{sales,qual}$ | Gumbel | 1.2165 | 0 |
| | $c_{sales,grad}$ | SClayton | 0.4361 | 0 |
| | $c_{sales,rise2}$ | SClayton | 0.1089 | 0 |
| | $c_{sales,emp}$ | Frank | 13.2271 | 0 |
| | $c_{sales,rise}$ | Gaussian | 0.5924 | 0 |
| | $c_{sales,prom}$ | Gumbel | 1.1635 | 0 |
| | $c_{sales,horiz}$ | SClayton | 0.1520 | 0 |
| 2 | $c_{ext,qual \mid sales}$ | Gaussian | 0.0404 | 0 |
| | $c_{ext,grad \mid sales}$ | Student t | 0.6526 | 2 |
| | $c_{ext,rise2 \mid sales}$ | SJoe | 1.9468 | 0 |
| | $c_{ext,emp \mid sales}$ | SJoe | 1.1277 | 0 |
| | $c_{ext,rise \mid sales}$ | Student t | 0.3288 | 2 |
| | $c_{ext,prom \mid sales}$ | Student t | 0.8141 | 2 |
| | $c_{ext,horiz \mid sales}$ | SJoe | 4.2206 | 0 |
| 3 | $c_{qual,rise \mid sales,ext}$ | Independent | 0 | 0 |
| | $c_{qual,prom \mid sales,ext}$ | Independent | 0 | 0 |
| | $c_{qual,horiz \mid sales,ext}$ | SJoe | 1.0001 | 0 |
| | $c_{qual,grad \mid sales,ext}$ | Gaussian | 0.1799 | 0 |
| | $c_{qual,rise2 \mid sales,ext}$ | Frank | 0.7868 | 0 |
| | $c_{qual,emp \mid sales,ext}$ | Independent | 0 | 0 |
| 4 | $c_{grad,horiz \mid sales,ext,qual}$ | Independent | 0 | 0 |
| | $c_{grad,emp \mid sales,ext,qual}$ | SJoe | 1.0001 | 0 |
| | $c_{grad,rise \mid sales,ext,qual}$ | Independent | 0 | 0 |
| | $c_{grad,rise2 \mid sales,ext,qual}$ | Frank | 0.8799 | 0 |
| | $c_{grad,prom \mid sales,ext,qual}$ | Frank | 2.7894 | 0 |
| 5 | $c_{rise2,prom \mid sales,ext,qual,grad}$ | Gaussian | -0.2264 | 0 |
| | $c_{rise2,horiz \mid sales,ext,qual,grad}$ | Joe | 1.5419 | 0 |
| | $c_{rise2,emp \mid sales,ext,qual,grad}$ | Independent | 0 | 0 |
| | $c_{rise2,rise \mid sales,ext,qual,grad}$ | Independent | 0 | 0 |
| 6 | $c_{emp,prom \mid sales,ext,qual,grad,rise2}$ | Independent | 0 | 0 |
| | $c_{emp,horiz \mid sales,ext,qual,grad,rise2}$ | Independent | 0 | 0 |
| | $c_{emp,rise \mid sales,ext,qual,grad,rise2}$ | Independent | 0 | 0 |
| 7 | $c_{rise,prom \mid sales,ext,qual,grad,rise2,emp}$ | Independent | 0 | 0 |
| | $c_{rise,horiz \mid sales,ext,qual,grad,rise2,emp}$ | Independent | 0 | 0 |
| 8 | $c_{prom,horiz \mid sales,ext,qual,grad,rise2,emp,rise}$ | Independent | 0 | 0 |

Table 1: Assolombarda data: selected copulas and C-vine PCC parameters. SClayton and SJoe denote the Survival Clayton and Survival Joe copula, respectively.

| **FTSE-MIB data: Parameters of the C-Vine** | | | | |
|---|---|---|---|---|
| *Tree* | *Copulas* | *Family* | *Parameter 1* | *Parameter 2* |
| 1 | $c_{sales,ncliab}$ | Frank | 4.7259 | 0 |
| | $c_{sales,stocks}$ | Frank | 7.6876 | 0 |
| | $c_{sales,ncas}$ | Joe | 3.5353 | 0 |
| | $c_{sales,goodwill}$ | Frank | 4.4699 | 0 |
| | $c_{sales,emp}$ | Joe | 3.3784 | 0 |
| | $c_{sales,prov}$ | Gaussian | 0.8769 | 0 |
| | $c_{sales,cliab}$ | Student t | 0.5715 | 2 |
| 2 | $c_{ncliab,ncas \mid sales}$ | Independent | 0 | 0 |
| | $c_{ncliab,goodwill \mid sales}$ | Gumbel | 1.0211 | 0 |
| | $c_{ncliab,emp \mid sales}$ | Frank | 4.1546 | 0 |
| | $c_{ncliab,prov \mid sales}$ | Gaussian | -0.0846 | 0 |
| | $c_{ncliab,cliab \mid sales}$ | Frank | 1.8065 | 0 |
| | $c_{ncliab,stocks \mid sales}$ | Clayton | 0.4768 | 0 |
| 3 | $c_{stocks,ncas \mid sales,ncliab}$ | Frank | -1.6267 | 0 |
| | $c_{stocks,emp \mid sales,ncliab}$ | Independent | 0 | 0 |
| | $c_{stocks,prov \mid sales,ncliab}$ | Independent | 0 | 0 |
| | $c_{stocks,cliab \mid sales,ncliab}$ | Independent | 0 | 0 |
| | $c_{stocks,goodwill \mid sales,ncliab}$ | Frank | -0.0655 | 0 |
| 4 | $c_{ncas,goodwill \mid sales,ncliab,stocks}$ | Independent | 0 | 0 |
| | $c_{ncas,emp \mid sales,ncliab,stocks}$ | Independent | 0 | 0 |
| | $c_{ncas,prov \mid sales,ncliab,stocks}$ | Frank | 1.9324 | 0 |
| | $c_{ncas,cliab \mid sales,ncliab,stocks}$ | Independent | 0 | 0 |
| 5 | $c_{goodwill,emp \mid sales,ncliab,stocks,ncas}$ | Independent | 0 | 0 |
| | $c_{goodwill,prov \mid sales,ncliab,stocks,ncas}$ | Independent | 0 | 0 |
| | $c_{goodwill,cliab \mid sales,ncliab,stocks,ncas}$ | Independent | 0 | 0 |
| 6 | $c_{emp,cliab \mid sales,ncliab,stocks,ncas,goodwill}$ | Independent | 0 | 0 |
| | $c_{emp,prov \mid sales,ncliab,stocks,ncas,goodwill}$ | Independent | 0 | 0 |
| 7 | $c_{prov,cliab \mid sales,ncliab,stocks,ncas,goodwill,emp}$ | Independent | 0 | 0 |

Table 2: FTSE-MIB data: selected copulas and C-vine PCC parameters.

| | **Unconditionalized NPBBN** | | | **Predictive reasoning** | | | **Diagnostic reasoning** | | |
|---|---|---|---|---|---|---|---|---|---|
| *Node name* | *Median* | *Mean* | *Std Dev* | *Median* | *Mean* | *Std Dev* | *Median* | *Mean* | *Std Dev* |
| sales | 4602162 | 12595630 | 30800230 | 122869 | 194510 | 238424 | 188000 | 188000 | 0 |
| cliab | 179254 | 353552300 | 32324790000 | 23141 | 31746140 | 2511286000 | 10199 | 6047388 | 141801900 |
| prov | 241824 | 1648392 | 8959909 | 9632 | 30919 | 91406 | 17948 | 59557 | 179661 |
| employees | 13058 | 37930 | 95616 | 365 | 365 | 0 | 1497 | 2786 | 4194 |
| goodwill | 213450 | 991221100 | 40223790000 | 646 | 327970 | 13146760 | 9034 | 15427590 | 701917500 |
| ncas | 116541 | 33019030 | 1172014000 | 50 | 50 | 0 | 1799 | 83431 | 1016132 |
| stocks | 264140 | 1741744 | 7913491 | 240 | 240 | 0 | 20127 | 69202 | 201553 |
| ncliab | 751678 | 297541300 | 7431751000 | 4082 | 663782 | 15698810 | 15993 | 1947262 | 44543930 |

Table 3: Predictive and diagnostic reasoning with the NPBNN for the FTSE-MIB data. The node names are in the first columns. Columns 2–4 list, respectively, the medians, means and standard deviations for the unconditionalized NPBBN; columns 5–7 list the same results for the case of predictive reasoning (for fixed low values of "emp", "ncas" and "stocks"); columns 8–10 list the same results for the case of diagnostic reasoning (for a fixed low value of "sales").

| **Assolombarda data simulation study** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **t Copula** | | | **C-Vine** | | | **NPBNN** | | |
| *Node name* | *Median* | *Mean* | *Std Dev* | *Median* | *Mean* | *Std Dev* | *Median* | *Mean* | *Std Dev* |
| sales | 0.92400 | 0.82845 | 0.19808 | 0.89167 | 0.81490 | 0.19948 | 0.92400 | 0.83940 | 0.21473 |
| emp | 0.92400 | 0.82596 | 0.19879 | 0.85934 | 0.79710 | 0.22671 | 0.92400 | 0.83345 | 0.19362 |
| rise | 0.96769 | 0.91118 | 0.16251 | 0.99051 | 0.89730 | 0.16023 | 0.96769 | 0.89625 | 0.16639 |
| rise2 | 0.96769 | 0.89204 | 0.16387 | 0.96769 | 0.90204 | 0.15977 | 0.99051 | 0.92632 | 0.12693 |
| prom | 0.99847 | 0.92378 | 0.15118 | 0.99051 | 0.91868 | 0.14088 | 0.99847 | 0.93451 | 0.13280 |
| horiz | 0.99991 | 0.96869 | 0.08892 | 0.99991 | 0.96431 | 0.09411 | 0.99991 | 0.96999 | 0.08076 |
| ext | 0.38623 | 0.40424 | 0.21919 | 0.34882 | 0.35135 | 0.21739 | 0.42365 | 0.40504 | 0.24651 |
| grad | 0.99847 | 0.95135 | 0.10131 | 0.99991 | 0.95047 | 0.11797 | 0.99847 | 0.94842 | 0.11772 |
| qual | 1.00000 | 0.96419 | 0.08990 | 1.00000 | 0.97651 | 0.06374 | 1.00000 | 0.99415 | 0.02427 |

Table 4: Simulation study for the Assolombarda data. The node names are in the first columns. Columns 2–4 list, respectively, the medians, means and standard deviations of the Kolmogorov-Smirnov p-values for the multivariate t copula; columns 5–7 list the same results for the C-vine; columns 8–10 list the same results for the NPBBN.

| **FTSE-MIB data simulation study** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **t Copula** | | | **C-Vine** | | | **NPBNN** | | |
| *Node name* | *Median* | *Mean* | *Std Dev* | *Median* | *Mean* | *Std Dev* | *Median* | *Mean* | *Std Dev* |
| sales | 0.95144 | 0.82924 | 0.19020 | 0.95144 | 0.87035 | 0.18840 | 0.95144 | 0.83460 | 0.21944 |
| cliab | 0.95144 | 0.83322 | 0.19926 | 0.95144 | 0.82044 | 0.22950 | 0.95144 | 0.83888 | 0.21495 |
| prov | 0.94549 | 0.82947 | 0.20143 | 0.86366 | 0.78658 | 0.22774 | 0.86366 | 0.81087 | 0.20438 |
| employees | 0.95144 | 0.81756 | 0.21390 | 0.95144 | 0.82977 | 0.21854 | 0.95144 | 0.81719 | 0.23339 |
| goodwill | 0.95144 | 0.81703 | 0.21838 | 0.95144 | 0.85069 | 0.18576 | 0.95144 | 0.85779 | 0.18791 |
| ncas | 0.78183 | 0.81139 | 0.18362 | 0.94549 | 0.83495 | 0.20541 | 0.94549 | 0.83036 | 0.20573 |
| stocks | 0.79123 | 0.80044 | 0.21316 | 0.95144 | 0.82966 | 0.20268 | 0.95144 | 0.85187 | 0.18010 |
| ncliab | 0.94549 | 0.82689 | 0.19431 | 0.94549 | 0.86051 | 0.17019 | 0.94549 | 0.86351 | 0.19108 |

Table 5: Simulation study for the FTSE-MIB data. The node names are in the first columns. Columns 2–4 list, respectively, the medians, means and standard deviations of the Kolmogorov-Smirnov p-values for the multivariate t copula; columns 5–7 list the same results for the C-vine; columns 8–10 list the same results for the NPBBN.