Official Statistics Data Integration Using Copulas

Luciana Dalla Valle

Plymouth University

April 11, 2014

2014 ENBIS-SFds Spring Meeting
Summary

- **Aim**: integrate financial information, incorporating the dependence structure among the variables.
- **Methodology**: two types of graphical models, based on copulas
  - **Vines**: undirected graphs, representing pair copula constructions, which are used to model the dependence structure of a set of variables.
  - **Non parametric Bayesian belief nets (NPBBNs)**: directed graphs, that use pair copulas to model the dependencies, and allow for diagnosis and prediction via conditionalization.
- **Application**: two financial datasets
  - **Assolombarda dataset**: data collected through a survey
  - **FTSE-MIB dataset**: official statistics data.
Advances in technology and communications have increased the availability of sources of information and **large databases**.

**Multivariate modeling** is of fundamental interest and new methods to manipulate high quantities of data have become essential.

**Data integration** has become an important issue due to the growth of the number of available data sources and to the increase in data quality standards.

It is fundamental to **integrate** information obtained from **specific datasets** with those obtained from **official statistics**.
Existing literature about data integration:

- **Multivariate regression**: Foresti et al. (2012) used OLS to identify the determinants of sales growth, applying it to several integrated private databases.
  → *Cannot capture complex multivariate dependencies.*

- **Probabilistic graphical models**: represent multivariate densities via a combination of a qualitative graph structure that encodes independencies and local quantitative parameters.
  → *These models are limited to the discrete or normal cases.*
Motivations: significant departures from normality and complex dependence structures.

The word *Copula* is derived from Latin, meaning to bind, tie, connect.

The copula is a multivariate distribution function with marginals distributed according to a uniform on the interval $[0, 1]$.

This function, once applied to the univariate marginal distributions, returns their joint multivariate distribution, enclosing all the information about the dependence structure of the marginals.

The copula expresses the dependence structure of a set of random variables, whatever is the distribution of these variables (marginals).
The definition of Copula

Consider $X_1, \ldots, X_d$ to be random variables and $F$ their joint distribution function. Then we have the following definition.

**Definition: Copula**

The Copula associated with $F$ is a distribution function, $C : [0, 1]^d \rightarrow [0, 1]$, of random variables $X_1, \ldots, X_d$ with standard uniform marginal distributions $F_1, \ldots, F_d$ with the following properties:

1. $\forall (u_1, \ldots, u_d) \in [0, 1]^d$, then $C(u_1, \ldots, u_d) = 0$ if at least one coordinate of $(u_1, \ldots, u_d)$ is 0;

2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$, for all $u_i \in [0, 1], (i = 1, \ldots, d)$.

Hence, if $C$ is a Copula, then it is the distribution of a multivariate uniform random vector.
Sklar’s theorem

Sklar’s theorem (Sklar, 1959)

Let $F$ denote a $d$-dimensional distribution function with margins $F_1, \ldots, F_d$. Then there exists an $d$-copula $C$ such that for all $(x_1, \ldots, x_d)$

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

If $F_1, \ldots, F_d$ are all continuous, then the copula is unique; otherwise $C$ is uniquely determined. Conversely, if $C$ is a copula and $F_1, \ldots, F_d$ are distribution functions, then the function $F$ is a joint distribution with margins $F_1, \ldots, F_d$.
Copula density

- The joint density function is

\[
f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d)
\]

where \( c(F_1(x_1), \ldots, F_d(x_d)) \) is the \( d \)-variate copula density:

\[
c(F_1(x_1), \ldots, F_d(x_d)) = \frac{\partial^d C(F_1(x_1), \ldots, F_d(x_d))}{\partial F_1(x_1) \cdots \partial F_d(x_d)}.
\]

- Bivariate case \((d = 2)\):

\[
f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1)f_2(x_2)
\]

\[
f_{2|1}(x_2|x_1) = c_{12}(F_1(x_1), F_2(x_2)) \cdot f_2(x_2)
\]
Common bivariate copulas

**Elliptical copulas**
- Construction through inversion of Sklar’s theorem:

\[ C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)), \quad u_1, u_2 \in (0, 1) \]

where \( F \) is elliptical.
- **Gaussian** (from bivariate normal distribution with correlation \( \rho \)).
- **Student's t** (from bivariate Student's t distribution with \( \nu \) degrees of freedom and association \( \rho \)).

**Archimedean copulas**
- Construction through generator \( \varphi \) (McNeil and Neslehova 2009):

\[ C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2)), \quad u_1, u_2 \in (0, 1) \]

- **Clayton, Gumbel, Frank, ...**
Copula contour plots

Bivariate contour plots of different copulæ, with standard normal margins and $\tau = 0.5$
Pair Copula Constructions (PCCs) motivation

- The existing literature on copulas mainly focuses on the bivariate case.
- In the multivariate case, Gaussian and Student’s t copula are often not flexible enough to represent complex dependence structure of financial data.
- These multivariate extensions imply additional restrictions on the parameters that limit their flexibility.

Therefore…

⇒ Use PCCs
**Pair Copula Constructions**

- **PCCs** were originally proposed by Joe (1996), and later discussed in detail by Bedford and Cooke (2001 and 2002), Kurowicka and Cooke (2006), Aas et al. (2009) and Czado (2010).
- A PCC represents the complex pattern of dependence of multivariate data via a **cascade of bivariate copulas**.
- **PCCs** allow to construct **flexible** high-dimensional copulas by using only **bivariate** copulas as building blocks.
Pair Copula Construction in 3 dimension

Factorization

\[ f(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_1(x_1) \]

Using Sklar’s Theorem for \( f(x_1, x_2) \), \( f_{13|2}(x_1, x_3|x_2) \) and \( f(x_2, x_3) \) implies

\[ f_{2|1}(x_2|x_1) = c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) \]

\[ f_{3|12}(x_3|x_1, x_2) = f_{13|2}(x_1, x_3|x_2) \frac{1}{f_{1|2}(x_1|x_2)} \]

\[ = c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) f_{1|2}(x_1|x_2) f_{3|2}(x_3|x_2) \frac{1}{f_{1|2}(x_1|x_2)} \]

\[ = c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \]

3-dimensional PCC

\[ f(x_1, x_2, x_3) = c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \]

\[ \times c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) f_1(x_1) \]
Pair Copula Construction in $d$ dimension

\[ f(x_1, \ldots, x_d) = \prod_{\tau=1}^{d} f_{\tau}(x_{\tau}) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}, \]

where:

- $f_{\tau}$: $d$ marginal densities
- $c_{i,i+j|i+1,\ldots,i+j-1}(F(x_i|x_{i+1}, \ldots, x_{i+j-1}), F(x_{i+j}|x_{i+1}, \ldots, x_{i+j-1}))$: bivariate copulas
- $F(\cdot|\cdot)$: conditional distribution functions.
Regular Vines

Bedford and Cooke (2001, 2002) introduced a graphical model called Regular Vine to organize PCCs.

Definition: Regular vine (R-vine)

A Regular Vine on $d$ variables is a set of connected trees $T_1, \ldots, T_{d-1}$ with nodes $N_i$ and edges $E_i$ (for $i = 1, \ldots, d - 1$) satisfying

1. $T_1$ has nodes $N_i = \{1, \ldots, d\}$ and edges $E_1$;
2. for $i = 2, \ldots, d - 1$ the tree $T_i$ has nodes $N_i = E_{i-1}$;
3. two edges in tree $T_i$ are joined in tree $T_{i+1}$ if they share a common node in tree $T_i$. 
Simple Example: $d = 3$
Example: R-Vine
Example: Drawable vine (D-vine)

A D-vine is a regular vine where all nodes do not have degree higher than 2, that is each node is connected to no more than two other nodes.
**Example: Canonical Vine**

**Example: Canonical Vine (C-Vine)**
An R-Vine is called a canonical vine (C-vine) if each tree is a star and has a unique root node.
Regular vine distributions and copulas

Regular vine distribution

A \textit{d}-dimensional regular vine distribution has the following components

1. a \textit{regular vine tree} structure;
2. each edge corresponds to a \textit{pair copula} density;
3. the density of a regular vine distribution is defined by
   - the product of \textit{pair copula densities} over the \(d(d-1)/2\) edges identified by the regular vine trees
   - the product of the \textit{marginal} densities.
Example of Regular vine distribution

Simple Example: $d = 3$

A regular vine copula is defined as the product of pair copulas determined through the regular vine.

Example:

$$f(x_1, x_2, x_3) = c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))c_{23}(F_2(x_2), F_3(x_3))$$

$$\times c_{12}(F_1(x_1), F_2(x_2))f_3(x_3)f_2(x_2)f_1(x_1)$$
The proposed approach: theoretical framework

Regular vine estimation

Specification of:

- **Vine structure**
  - Choice amongst **C-vine, D-vine, R-vine, ...**
  - **maximal spanning tree algorithm**: capture the strongest dependencies in the first tree and to obtain a *parsimonious* model.

- **Copula families**
  - A copula family for each pair of variables selected using **Akaike Information Criterion (AIC)**.
  - Choice amongst: **elliptical** copulas (Gaussian and Student's t) and **archimedean** copulas (Clayton, Gumbel, Frank, Joe, and their rotated versions).

- **Copula parameters**
  - expressing *dependencies*, estimated using the **maximum likelihood method** (Aas et al. (2009)).
Conditional independence

- **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 (2007)) is performed on each pair of variables.

![Diagram](image)

**Figure**: Full (left) and simplified (right) D-vine structures
NPBBNs: Background

- **R-vines** have been successfully applied to datasets with dimensionality of at most tens of variables (Brechmann and Czado (2013)). However, with datasets of dimensionality of hundreds of variables, R-vines become **computationally intractable**.

- **Bayesian belief nets (BBNs)** are directed acyclic graphs (DAGs) whose **nodes** represent variables and the **arcs** represent causal relationships between the variables.

- The most popular classes of BBNs are **discrete** or **normal**. However, their **limitations** are
  - **discrete** BBNs are only suitable to datasets of limited size and complexity,
  - **normal** BBNs are limited by the assumption of joint normality.

- To overcome this limitations, Kurowicka and Cooke in 2006 introduced **NPBBNs**, where distributions can conform to **any parametric form** and the relationships among variables are defined through **R-vines**.
The proposed approach: theoretical framework

NPBBNs

- The direct predecessors of a node, corresponding to a variable, are called **parents**, while the direct descendants of a node are called **children**.
- The conditional independence statements encoded in the graph allow us to write the **joint density** as

\[
f(x_1, \ldots, x_d) = \prod_{j=1}^{d} f_{x_j|\text{Pa}(j)}(x_j|\text{Pa}_j)
\]

where
- \(f_{x_j|\text{Pa}(j)}\): conditional probability function associated to node \(j\), that corresponds to variable \(X_j \ (j = 1, \ldots, d)\)
- \(\text{Pa}_j\): set of all \(j\)'s parents.

- The **nodes** are associated with continuous invertible distributions, while each **arc** is represented by a **conditional rank correlation** between parent and child.
The proposed approach: theoretical framework
Continuous Non Parametric Bayesian Belief Nets

Example: NPBBNs

**Example:** A NPBBNs on four variables with conditional rank correlations assigned to arcs.

Assignments for the **DAG of the NPBBN**:

1. Construct a sampling **order** of the nodes and index the nodes according to it. We choose, i.e. the ordering (1, 2, 3, 4);

2. **Factorize the joint** following the sampling order, highlighting the **nodes** in each conditioning set that are not **parents** of the conditioned variable:

   \[ f(x_1, x_2, x_3, x_4) = f(1)f(2|1)f(3|12)f(4|321); \]

3. The **rank correlations** to be assigned to the arcs are \( \{ r_{12}, r_{13}, r_{43}, r_{42|3} \} \).
NPBBNs: Theorem

**Theorem**

Given:

- a directed acyclic graph with $d$ nodes specifying conditional independence relationships in a BBN;
- $d$ variables, assigned to the nodes, with continuous invertible distribution functions;
- the specification of conditional rank correlations on the arcs of the BBN;
- a copula realizing all correlations [-1, 1] for which correlation 0 entails independence;

the **joint distribution** of the $d$ variables is **uniquely determined**.

This joint distribution satisfies the characteristic factorization of the BBN and the conditional rank correlations are algebraically independent.
NPBBNs: sampling

- No analytical/parametric form of the joint distributions is available. Therefore, we sample the NPBBN using the procedures for Vines.
- For each term of the factorization a Vine is built, whose (conditional) rank correlations exactly correspond to those of the NPBBN.
- The (conditional) rank correlations and the marginal distributions needed to completely specify the NPBBN can be retrieved from data or elicited from experts.
Assolombarda dataset

- **Assolombarda** 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.

- 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.
Assolombarda variables

The variables in the Assolombarda 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 $d = 9$. 
Canonical Vine

Since sales is the target variable and dominates the dependencies with all the remaining variables, we used a C-vine and we set sales as the root node.

Figure: First (left) and second (right) C-vine trees for the Assolombarda data.
Non Parametric Bayesian Belief Nets

Figure: NPBBN for the Assolombarda data. Variables are represented with nodes.
Conditionalized NPBBNs: predictive reasoning

Figure: Conditionalized NPBBN for the Assolombarda data. The NPBBN is conditionalized for a high value of emp (predictive reasoning).

→ All variables are right-skewed

Employees from 364 to 32,000 → Sales from 188,000 to 4,745,000
Conditionalized NPBBNs: diagnostic reasoning

**Figure**: Conditionalized NPBBN for the Assolombarda data. The NPBBN is conditionalized for a high value of sales (diagnostic reasoning).

*Sales from 188,000 to 700,000 → Employees from 364 to 1,076 and Employees in external market from 5 to 9*
FTSE-MIB dataset

- FTSE-MIB (formerly MIB30) data is 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, referring to 2007, 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 variables in the FTSE-MIB dataset 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.

Therefore, the **dimensionality** of the dataset is $d = 8$. 
Canonical Vine

As in the previous example, since sales is the target variable and dominates the dependencies of the whole dataset, we used a C-vine and we set sales as the root node.

Figure: First (left) and second (right) C-vine trees for the FTSE-MIB data.
Non Parametric Bayesian Belief Nets

Figure: NPBBN for the FTSE-MIB data. Variables are represented with nodes.
Conditionalized NPBBNs: predictive reasoning

Figure: Conditionalized NPBBN for the FTSE-MIB data. The NPBBN is conditionalized for low value of \textit{emp}, \textit{ncas} and \textit{stocks} (predictive reasoning).

\rightarrow \textbf{All variables are right-skewed}

\textbf{Employees=365, Stock=240, Non-current assets=50} \rightarrow \textbf{Sales=194,510}
Conditionalized NPBBNs: diagnostic reasoning

Figure: Conditionalized NPBBN for the FTSE-MIB data. The NPBBN is conditionalized for a low value of sales (diagnostic reasoning).

Sales=188,000 $\rightarrow$ Employees=2,786, Non-current assets=83,431 and Stocks=6,9202
Simulation study

- We generated **1000 simulations** of the two datasets using **C-vines** and **NPBBNs**, and we compared the distribution of the original variables with the simulated variables.
- We considered the **multivariate t copula** as a benchmark, standard choice for financial data.
- We performed the **Kolmogorov-Smirnov test** for the equality of distributions for each simulation and we calculated the p-values. The closer to 1 the better the fit.
- **Results**: C-vine and NPBBN perform better than the traditional multivariate t copula.
Conclusions

- We presented a new approach to integrate the information provided by official sources with information provided by other sources.
- We used Vines to model the dependence 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 for predictive and diagnostic reasoning.
- We calibrated the two datasets via conditionalization to see what characteristics a set of firms should have in order to perform similarly to the firms described in the official data source.
References


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.


