Modeling high-dimensional dependence with directed acyclic graphs

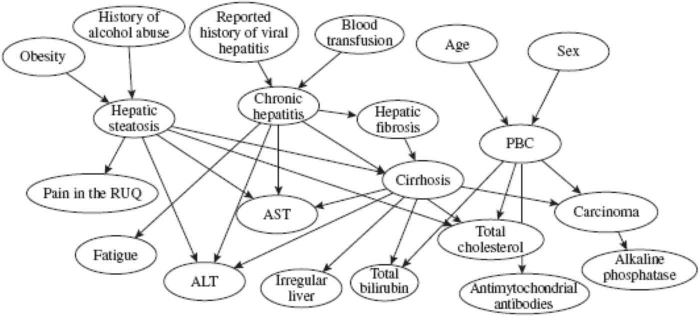
Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

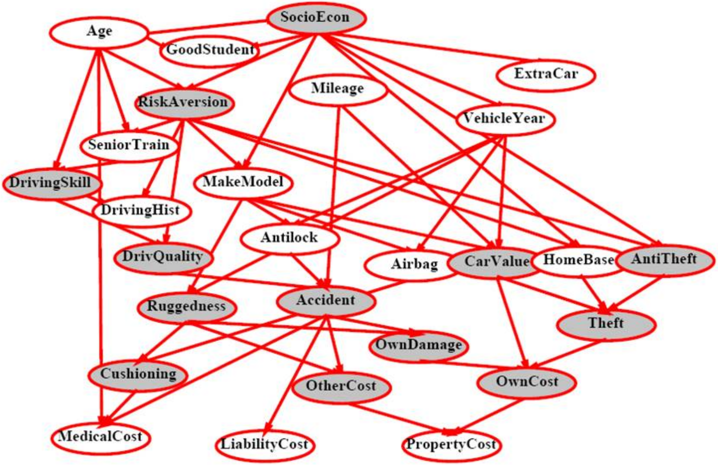
Directed acyclic graph

A directed acyclic graph (DAG), is a directed graph with no directed cycles. It has been widely used in many areas such as computer science, engineering, medical diagnosis, and crime risk factors analysis etc.

Example: Medical diagnosis



Example: Car insurance



Directed acyclic graph

Let us consider n random variables $X_1, \ldots, X_n$, a directed acyclic graph with n numbered nodes, and suppose node j ($1 \leq j \leq n$) of the graph is associated to the X_j variable. Then the graph is a Bayesian network, representing the variables $X_1, \ldots, X_n$, if:

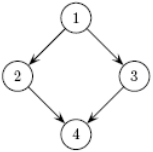
$$P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P\left(X_j \mid parents(X_j)\right),$$

or

$$f(x_1, \ldots, x_n) = \prod_{j=1}^{n} f\left(x_j \mid parents(x_j)\right),$$

where $parents(X_j)$ denotes the set of all variables X_i , such that there is an arc from node i to node j in the graph.

Example:



$$f(x_1, x_2, x_3, x_4) = f_1(x_1) f_{2|1}(x_2|x_1) f_{3|1}(x_3|x_1) f_{4|23}(x_4|x_2, x_3).$$

Any joint probability distribution may be represented by a Bayesian network.

Multivariate distribution

Multivariate real-valued distributions are of paramount importance in a variety of fields ranging from computational biology and neuro-science to economics to climatology. Choosing and estimating a useful form for the marginal distribution of each variable in the domain is often a straightforward task. In contrast, aside from the normal representation, few univariate distributions have a convenient multivariate generalization. Indeed, modeling and estimation of flexible (skewed, multi-modal, heavy tailed) high-dimensional distributions is still a formidable challenge.

In the literature, standard references on the theory of graphical models are mainly limited to the assumption of joint normality as far as continuous variables are concerned. At the same time, it is well known from the literature on statistical models for financial markets that the assumption of joint normality may lead to severe underestimation of certain risks and, in a more general sense, fails to yield suitable models in many applications, see, for instance, McNeil et al. (2005).

How to model high dimensional non-normal (or non-Gaussian) dependence?

Copula

Assume that $(X_1, \ldots, X_n)$ is a set of real-valued continuous random variables. Then, the joint distribution function is

$$F(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n).$$

Then, by Sklar Theorem, there exist a unique Copula function C such that

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)).$$

For the joint density function, we have,

$$f(x_1, \ldots, x_n) = c(F_1(x_1), \ldots, F_n(x_n)) \prod_{i=1}^{n} f_i(x_i),$$

where c is the copula probability density function. Or

$$C(u_1, \ldots, u_n) = F\left(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\right),$$

and

$$c(u_1, \ldots, u_n) = \frac{f\left(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\right)}{\prod_{i=1}^{n} f_i\left(F_i^{-1}(u_i)\right)}$$

Examples

1. Clayton (lower tail dependence)

$$C(u,v) = \left[\max\left\{u^{-\theta} + v^{-\theta} - 1, 0\right\}\right]^{-1/\theta}$$

2. Gumbel (upper tail dependence)

$$C(u,v) = \exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{1/\theta}\right\}$$

3. Gauss (no tail dependence)

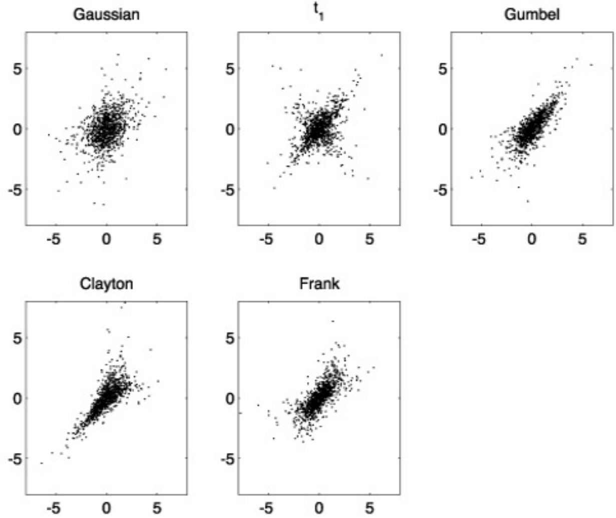
$$C(u,v) = \Phi\left(\Phi^{-1}(u), \Phi^{-1}(v)\right).$$

4. Product

$$C(u,v) = uv.$$

There are a good amount of copula functions, see, for example, Nelson (2006).

Copula



Copula-the emperor's new clothes?

THOMAS MIKOSCH (Extreme, 2005)

- There is no particular advantage of using copulas when dealing with multivariate distributions. Instead one can and should use any multivariate distribution which is suited to the problem at hand and which can be treated by statistical techniques.
- The marginal distributions and the copula of a multivariate distribution are inextricably linked. The main selling point of the copula technology -separation of the copula (dependence function) from the marginal distributions -leads to a biased view of stochastic dependence, in particular when one fits a model to the data.
- Various copula models (Archimedean, t-, Gaussian, elliptical, extreme value) are mostly chosen because they are mathematically convenient; the rationale for their applications is murky.
- Copulas are considered as an alternative to Gaussian models in a non-Gaussian world. Since copulas generate any distribution the class is too big to be understood and to be useful. There is little statistical theory for copulas. Sensitivity studies of estimation procedures and goodness-of-fit tests for copulas are unknown. It is unclear whether a good fit of the copula of the data yields a good fit to the distribution of the data.
- Copulas do not contribute to a better understanding of multivariate extremes.
- Copulas do not fit into the existing framework of stochastic processes and time series analysis; they are essentially static models and are not useful for modeling dependence through time.

No! At least the Emperor wears socks!

A note from Paul Embrechts-Journal of Risk and Insurance, 2009

Thomas Mikosch compared the copula craze with Hans Christian Andersen's fairy tale "The Emperor's new clothes" where the child says "But he hasn't got anything on!". In a recent publication, Kluppelberg and Resnick (2009), the authors end with "Religious Copularians have unshakable faith in the value of transforming a multivariate distribution to its copula. For the sceptics who believe the Emperor wears no clothes (Mikosch ,2005), perhaps use of the Pareto copula convinces some of these the Emperor at least wears socks." It is my personal belief that over the years to come, research will be able to put further garments on the poor man so that eventually in Hans Christian Andersen's words we can truly say "Goodness! How well they suit your Majesty! What a wonderful fit! What a cut! What colors! What sumptuous robes!".

Genest and Remillard-Extremes, 2006

Regretfully, he has chosen to produce a pamphlet which, written as it is in a lively but distinctly unscientific style, clearly does a disservice to the community by "throwing the baby out with the bathwater." While we respect his right to "ask some naive questions," we can hardly accept that either through ignorance or malice, he depicts copula theory as an unsubstantiated fad that leads to "a biased view of stochastic dependence."

Direct Decomposition

Based on

$$f(x_1, \ldots, x_n) = \prod_{j=1}^{n} f\left(x_j \mid parents(x_j)\right),$$

and

$$f(x_1, \ldots, x_n) = c(F_1(x_1), \ldots, F_n(x_n)) \prod_{i=1}^{n} f_i(x_i),$$

we can construct the following Bayesian model based on copulas:

$$f(x_1, \ldots, x_n) = \prod_{j=1}^{n} \frac{c(F_j(x_j), F_{parents(j)}(parents(x_j)))}{\int c(F_j(x_j), F_{parents(j)}(parents(x_j)))f_j(x_j)dx_j}.$$

Pair Decomposition

While there is a plethora of literature on bivariate copula families (also called pair-copula families), the range of higher-variate copula families is rather limited, see Joe (1997, Chapter 4). Many popular bivariate copulas have no straightforward multivariate extension.

Based on work of Joe (1996), Bedford and Cooke (2002, The Annals of Statistics) therefore proposed a flexible way of constructing multivariate copulas that uses (conditional) pair copulas as building blocks only. The core of their approach is a graphical representation called a regular vine that consists of a sequence of trees, each edge of which is associated with a certain pair copula. This idea was further developed by Aas et al. (2009, Insurance: Mathematics and Economics), and others.

Based on similar idea, Bauer, et al. (2011) developed the pair composition for Bayesian network.

Pair-copula constructions for Bayesian network

$D = (V, E)$ is a DAG with vertex set V and edge set E, and define

$$pa(v; w) := \{u \in pa(v) | u < w, w \in pa(v)\}.$$

For example, for $V = \{1, 2, 3, 4\}$, we have:

$$pa(1; \emptyset) = \emptyset, pa(2; 1) = \emptyset, pa(3; 1) = \emptyset, pa(4; 2) = \emptyset$$

$$pa(4; 3) = \{2\}.$$

For the Bayesian network with continuous distributions:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i) \prod_{w \in pa(v)} c_{vw}\left(F_{v|pa(v;w)}(x_v | x_{pa(v;w)}), F_{w|pa(v;w)}(x_w | x_{pa(v;w)}) | x_{pa(v;w)}\right).$$

In our example,

$$\begin{aligned} f(x_1, x_2, x_3, x_4) &= \prod_{i=1}^{4} f_i(x_i) \cdot c_{21}(F_2(x_2), F_1(x_1)) c_{31}(F_3(x_3), F_1(x_1)) c_{42}(F_4(x_4), F_2(x_2)) \\ &\times c_{43|2}(F_{4|2}(x_4|x_2), F_{3|2}(x_3|x_2)|x_2) \end{aligned}$$

MLE estimation

How can we derive the conditional copula, say, $c_{43|2}$? In practice, we just assume that they are unconditional. This assumption reduces model complexity while still encompassing a rich class of DAG copulas and has become common practice in likelihood inference for PCCs, see Aas et al. (2009) and Haff, et al. (2010).

The log-likelihood function in our example is

$$\begin{aligned}\log(L) &= \sum_{i=1}^{4} \log(f_i(x_i)) + \log(c_{21}(F_2(x_2), F_1(x_1))) + \log(c_{31}(F_3(x_3), F_1(x_1))) \\ &\quad + \log(c_{42}(F_4(x_4), F_2(x_2))) + \log(c_{43|2}(F_{4|2}(x_4|x_2), F_{3|2}(x_3|x_2)))\end{aligned}$$

Simplified pari-copula construction (PCC): simply useful or too simplistic?

What kinds of distributions can the simplified PCC represent?

Haff et al. (2010, JMVA) showed the following distributions:

1. Multivariate normal
2. Multivariate Pareto distribution of the forth kind
3. Multivariate Elliptical distributions

Not all multivariate distributions can be represented by a simplified PCC. However, as proved in Haff et al. (2010, JMVA), one can always use it as an approximation. In fact, it is a good one even though the simplifying assumption is far from being fulfilled.

Estimation

1. Marginal distributions
   1.1 Joe and Xu (1996): Inference functions for marginals (IFM)
   1.2 Genest et al. (1995): Transform univariate marginals to uniforms
2. Estimate parameters of Copulas

   After we estimate the marginal distributions, we select proper copulas to estimate their parameters. Note we estimate the parameters of each copula independently of the others.
3. Model selection

   BIC or AIC

Application: Financial returns

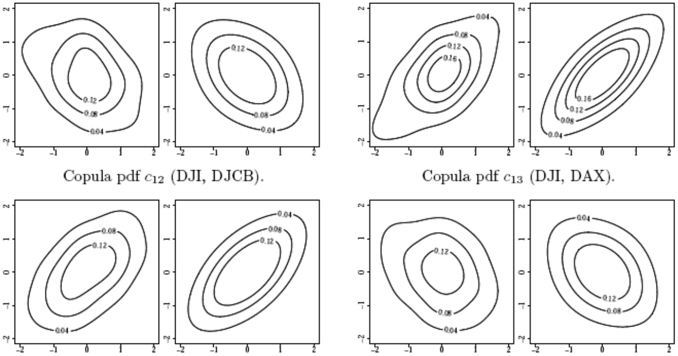
Purpose: Modeling dependence structure of daily log-returns of DJI, DJCB, DAX, RDAX.

DJI=Down Jones Industrial Average
DJCB=Dow Jones Corporate Bond Index
DAX=German stock index
RDAX=the corresponding German corporate bond index

Data: April 3rd, 2007 to Sep. 30, 2010.

Based on the economic consideration that the German stock index is driven by its US counterpart and that within the US and Germany corporate bond indices are driven by the respective national stock indices. That is
1=DJI
2=DJCB
3=DAX
4=RDAX

Choose proper copulas



Copula pdf c_{12} (DJI, DJCB).

Copula pdf c_{13} (DJI, DAX).

Copula pdf c_{24} (DJCB, RDAX).

Copula pdf $c_{34|2}$ ({DAX, RDAX} | DJCB).

Figure 6: Kernel density estimates of pair-copula pdfs (left) and our choices of Student's t copula pdfs (right) for modeling the DJI, DJCB, DAX, and RDAX data. All copulas are displayed with standard normal margins.

		Student C_{12}	Student C_{13}	Student C_{24}	Student $C_{34\|2}$	AIC
nG	S	$-0.35, 10.1$ $(0.07, 0.40)$	$0.66, 9.2$ $(0.11, 0.44)$	$0.57, 17.4$ $(0.09, 0.61)$	$-0.29, 8.9$ $(0.06, 0.36)$	-1002.0
	J	$-0.35, 10.4$ $(0.06, 0.35)$	$0.66, 9.3$ $(0.11, 0.40)$	$0.56, 14.0$ $(0.09, 0.51)$	$-0.29, 8.7$ $(0.07, 0.36)$	-1002.4
sG	S	$-0.35, 10.1$ $(0.07, 0.40)$	$0.66, 9.2$ $(0.11, 0.44)$	$0.57, 17.4$ $(0.09, 0.61)$	$-0.30, 9.4$ $(0.06, 0.36)$	-999.4
	J	$-0.35, 10.1$ $(0.07, 0.40)$	$0.66, 9.2$ $(0.11, 0.44)$	$0.56, 15.5$ $(0.08, 0.55)$	$-0.31, 9.1$ $(0.07, 0.44)$	-999.8
		Gauss C_{12}	Gauss C_{13}	Gauss C_{24}	Gauss $C_{34\|2}$	AIC
G	S	-0.34 (0.06)	0.66 (0.10)	0.57 (0.08)	-0.28 (0.06)	-971.5
	J	-0.34 (0.05)	0.66 (0.10)	0.57 (0.08)	-0.28 (0.06)	-971.5

Table 7: Sequential (S) and joint (J) ML estimates, standard errors (in parentheses), and AIC values for the Gaussian (G), the non-Gaussian (nG), and the semiparametric non-Gaussian (sG) DAG copula model applied to the log-return data.

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

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

A. Bauer, C. Czado and T. Klein, Pair-copula construction for non-Gaussian DAG models (2011).