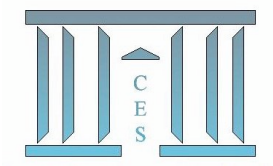
Note on New Prospects on Vines

Dominique Guegan, Pierre-André Maugis

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers



Note on New Prospects on Vines

Dominique GUEGAN, Pierre André MAUGIS

2008.95

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://ces.univ-paris1.fr/cesdp/CES-docs.htm
ISSN : 1955-611X

Note on New Prospects on Vines

D. Guégan*, P. A. Maugis†

Abstract

We present here a new way of building vine copulas that allows us to create a vast number of new vine copulas, allowing for more precise modeling in high dimensions. To deal with this great number of copulas we present a new efficient selection methodology using a lattice structure on the vine set. Our model allows for a lot of degrees of freedom, but further improvements face numerous problems caused by vines' complexity as an estimator in a statistical and computational way, problems that we will expose in this paper. Robust n-variate models would be a great breakthrough for asset and risk management in banks and insurance companies.

Keywords: Vines - Multivariate copulas - Model Selection

1 Introduction

For almost ten years now copulas have been used in econometrics and finance. It has became an essential tool for pricing complex products, managing portfolios and evaluating risk in banks: for instance to compute VaR (Value at Risk) and ES (Expected shortfall). Moreover copulas appear as a very flexible tool, allowing for semi-parametric estimation, fast parameter optimization and time varying parameters, making it a very interesting tool, but with one big lack: its use in high dimension.

Very recently Aas et al. (2007) produced a paper exposing a method to build copulas of any dimension using pair-copulas as building blocks. We call them vine copulas. The previous cited work uses partial conditioning to build multivariate models, as Joe (1997) and Bedford and Cooke (2002, 2001). Another possible way to build multivariate copulas is to use the "nested copulas". In this paper we focus on vine copulas, which appear richer than "nested copulas" for multivariate analysis (Berg and Aas, 2007).

*Paris School of Economics, CES-MSE, Université Paris 1 Panthéon-Sorbonne, 106 boulevard de l'Hopital 75647 Paris Cedex 13, France, e-mail: dguegan@univ-paris1.fr

†CES-MSE, Université Paris 1 Panthéon-Sorbonne, 106 boulevard de l'Hopital 75647 Paris Cedex 13, France, e-mail: p.a.maugis@gmail.com

Our purpose is to use vines efficiently to estimate n-variate densities. After recalling Aas et al. (2007) method and discussing some limitations of their approach, we provide a new point of view, focusing on the following points:

- The building of multivarriate copulas based on vines method permit to use any pair-copulas. We extend the works of Aas et al. (2007) and Berg and Aas (2007) who based their estimation procedure on Student pair-copulas and Gumbel-Hougaard pair-copulas.
- A vine is a decomposition of a n-variate density into a combination of bivariate densities. Such a decomposition is not unique, so that there exist many different vines. We present here new such decompositions allowing us to describe more varied n-variate densities.
- In order to select the best vine formula, we develop an efficient methodology using lattices. This method presents the advantage that it does not need to test all the models.

The paper is organized as follows. In Section 2, we give the definition of the vines. Section 3 provides some characteristics of vines which motivate our work. In Section 4, we introduce a new way to build vines and in Section 5 we describe the model selection procedure, relying on a lattice structure on the vine set. Section 6 concludes.

2 Vine Copula Definition

To build a vine copula we express a n-variate density function based on increasing conditionalities. Let us consider a vector $X = (X_1, X_2, \cdots, X_n)$ of n random variables with joint density function $f(x_1, \ldots, x_n)$ and cumulative density function $F(x_1, \ldots, x_n)$. In the following, we denote C the associated copulas and c their densities. We work in two steps. First we express the joint density as:

$$f(x_1, \ldots, x_n) = f(x_1).f(x_2|x_1).f(x_3|x_1, x_2) \cdots f(x_n|x_1, x_2, ..., x_{n-1}),$$

where the terms are respectively:

$$f(x_2|x_1) = c_{1,2}(F(x_1), F(x_2)).f(x_2),$$

and

$$f(x_3|x_1, x_2) = c_{1,2|3}(F(x_1|x_3), F(x_2|x_3)).f(x_1|x_3)$$

and so on. For instance, in dimension three, a possible vine formula is:

$$\begin{aligned} f(x_1, x_2, x_3) =& f(x_1).f(x_2).f(x_3) \\ & .c_{1,2}(F(x_1), F(x_2)).c_{1,3}(F(x_1), F(x_3)) \\ & .c_{1,3|2}(F(x_1|x_2), F(x_3|x_2)). \end{aligned}$$

In order to compute the cumulative density functions of the variables with conditionalities in a more general context, we use the following formula as many times as necessary, where ν is a set of variables and ν_j is one of them :

$$F(x|\nu) = \frac{\partial C_{x,\nu_j|\nu_{-j}}(F(x|\nu_{-j}), F(\nu_j|\nu_{-j}))}{\partial F(\nu_j|\nu_{-j})}, \tag{1}$$

Joe (1997). There are $|\nu|$ ways to compute (1) and taking into account the permutations of the variables, this means that there are about $\prod_{i=1}^{n} i!$ vine formulas. Aas et al. (2007) keeps only the "regular vines", reducing the possible deconstruction to $n!$. The regular vines have the characteristic of needing the smallest number of pair-copulas to build the model (in dimension n: $n(n-1)/2$ copulas). An explicit closed formula for these regular vines is obtained through two sub categories of copulas:

(i) The Canonical Vines:
$\prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j,j+1|1\dots j-1}\{F(x_j|x_1 \dots x_{j-1}), F(x_{j+1}|x_1 \dots x_{j-1})\}$,

(ii) The D-vines:
$\prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1\dots i+j-1}\{F(x_i|x_{i+1} \dots x_{i+j-1}), F(x_{i+j}|x_{i+1} \dots x_{i+j-1})\}$.

To estimate each formula, Aas et al. (2007) use a n-stage maximum likelihood. The simple pair-copulas are first evaluated. Then, they estimate those with one conditional variable and so on. The convergence of the maximum likelihood estimator is based on the fact that it is sufficient to efficiently estimate bivariate dependences to efficiently estimate the whole model, Bedford and Cooke (2002).

3 The Vine Estimator's Slip

Above we explained that the estimation of a vine formula is done by using (1) multiple times. We show here how the lack of most classical pair-copulas in the center of the distribution leads to erroneous estimation at the second stage of this procedure. And, with an example, how it leads to each vine formulas describing completely different behaviors.

3.1 Pair-copulas

Our model uses pair-copulas as building blocks. To model more varied relationships between two time-series we want to use varied pair-copulas[1]. Unfortunately most pair-copulas focus on the tail of the distribution, not on its center,

[1]To do this we need an effective pair-copulas selection procedure, until now no consensus exists on such a procedure. We will not address this question here, but the reader should keep it in mind.

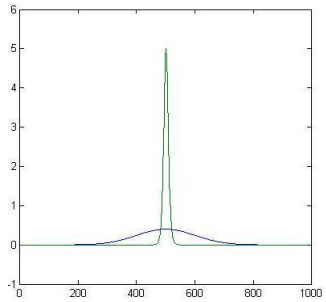


Figure 1: $f(X|Y = 0.5)$ computed with (1): $X, Y \sim U(0,1)$; $C_\theta(F(X), F(Y))$ Gumbel Copula ($\theta = 50$).

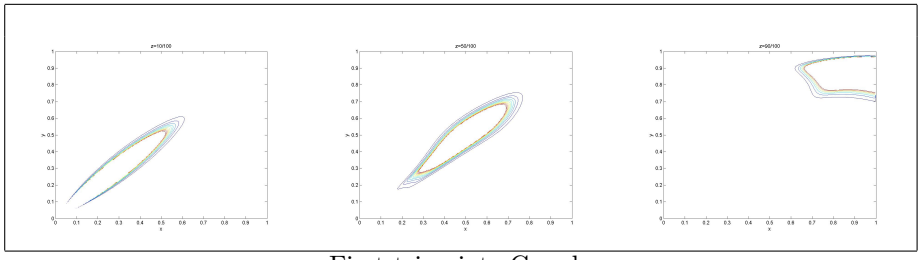
and this leads in our case to erroneous estimation; we illustrate this fact with a short example.

Let X, Y and Z be three random variables such that: $X, Y, Z \sim U(0,1)$ and that $C_\theta(F(X), F(Y))$ is a Gumbel copula of tail parameter $\theta = 50$. We are interested in computing $c(X, Y, Z)$. We represent $f(X|Y = 0.5)$ computed using (1) on Figure 1. We observe a strong concentration around the mean. Yet we do not expect such a behavior: the random variables should be nearly independent - as $Y = 0.5$ - so that $(X|Y = 0.5)$ should be close to a $U(0,1)$, which is not the case. In our example the copula greatly overstates the correlation in the center of the distribution, but what is more important are the consequences of this fact. At the next step we need to estimate the copula $C_{\theta'}(F(X|Y), F(Z|Y))$ assuming that its marginal distribution is similar to the one given in Figure 1. This means that $C_{\theta'}$ will be evaluated almost only on extreme events, making the estimation erroneous. Furthermore, the estimation of $C_{\theta'}$ will increase the tail parameters because the co-movements in the tails are heavily weighted. So that the third step of the estimation procedure will be even worst, and so on.

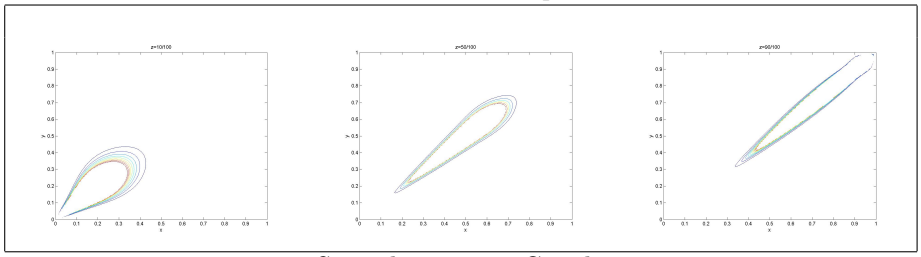
It appears with this example that the inability of most pair-copulas to capture distribution centers' leads, in the case of vines, to erroneous estimation of the whole distribution.

3.2 Example

Vines are strange statistical objects: they are decompositions of the same density and are as such theoretically equivalent. Nevertheless each evaluated formula can describe widely different behaviors. If each pair-copula was the true copula, this problem would not appear, but since we only have estimators, a vine corresponds to the model resulting from a scaffolding of different pairwise relationships between the variables. The following example shows how different two vine copulas can be even if they are estimated on the same data-set.



First trivariate Copula



Second trivariate Copula

Figure 2: Two copulas estimated on the same data set: slices at z = 10%, 50%, 90%

We evaluated two different trivariate vine copulas formulas on the same data set and then we compared them graphically. The data comes from Datastream, and is the daily Morgan Stanley evaluation of the French, German and British price indexes from the 1st of January 2006 to 1st of December 2007. Marginals are modelized using a $GARCH$ process with Student errors. We select pair-copulas using the AIC criteria, comparing Gumbel, Clayton and Gaussian copulas. Computations are done in $MatLab$. In Figure 2 we plot the estimated copulas with uniform margins.

The two trivariate copulas in Figure 2 have critically different tail behaviors. According to the first vine the three assets go down together, according to the second, they go up together. This implies different managing strategies for the manager who has in charge the portfolio composed by these three assets.

The lack of robustness for the pair-copulas creates for each estimated vines formula different behaviors. To compensate for this lack of robustness, we propose to use a great number of decompositions of the n-variate vector, so that the a "good" estimation of the pair-copulas will have more possibilities to occur. To discriminate between the good and the bad estimations, we will then present a model selection approach that can deal with the size of this set[2].

[2]This model selection approach is not data-snooping: the inability of classic copula to capture some dependence necessary to build a vine makes that estimated vines formulas have very varied behaviors, but all the vines are estimators of the same density. We are in fact always estimating the same model, but in different ways.

4 New Possible Vines

In this section, we propose a procedure allowing to build more possible vine formulas. Our procedure is less simple than Aas et al. (2007) approach and require greater computation time. Nevertheless, in fine, it will provide a more precise modeling. We first present a new algorithm to build vines and specify its properties. Then, we explicit the estimation procedure.

4.1 Formula

Let us consider a vector $X = (X_1, X_2, \cdots, X_n)$ of random variables with joint density function $f(x_1, \ldots, x_n)$. This density f can be factorised as follows:

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1, \ldots, x_{n-1}).f(x_n|x_1, \ldots, x_{n-1}) \\ &= f(x_1, \ldots, x_{n-1}).f((x_n, x_{n-1})|(x_{n-1}|x_1, \ldots, x_{n-2})) \\ &= f(x_1, \ldots, x_{n-1}).f(x_1, \ldots, x_{n-2}, x_n)/f(x_1, \ldots, x_{n-2}) \\ &\quad .c_{n,n-1|1\ldots n-2}(F(x_n|x_1 \ldots x_{n-2}), F(x_{n-1}|x_1 \ldots x_{n-2})). \end{aligned} \tag{2}$$

This formula allows us to compute a n-density with two (n-1)- and one (n-2)-densities. To use only efficient models, we need to simplify the denominator[3]. We will not explain further how this algorithm unwinds here (Guégan and Maugis, 2008), as it is a purely computational issue on how the vine formula is built, and is not relevant for our present purpose. Notice that it allows to build all possible vines but unwinds in a complex way.

This new formula (2) allows for more possible vine formulas. We first explain why the regular vines require the smallest number of pair-copulas and why the un-regular vines, even with more parameters, are still interesting as formulas.

4.2 Regular Vines Formulas

We will now build the formula with the smallest number of pair-copulas. We describe the first step of the algorithm and then indicate how the algorithm works at each step in order to minimize the number of pair-copulas[4]. This equation procedure generates regular vines.

[3] If we allowed for these denominators, divergence around zero would occur, and the convergence would not be as fast. Also such a simplification can only occur on the next step of the procedure, and not later, because otherwise the dimensions could not match.

[4] This method is sufficient to find the estimator with the smallest number of pair-copulas as we will reduce the number of distributions to be computed by matching them to those appearing at the next step, matching that can occur only then because otherwise the dimensions of the distributions could not match.

- Let us first compute $F(x_n|x_1 \dots x_{n-2})$ and $F(x_{n-1}|x_1 \dots x_{n-2})$ using the relationship (1) for any variable j:

$$\frac{\partial C_{x_n,x_j|x_1,\dots,x_j,\dots,x_{n-2}}(F(x_n|x_1,\dots,x_j,\dots,x_{n-2}),F(x_j|x_1,\dots,x_j,\dots,x_{n-2}))}{\partial F(x_j|x_1,\dots,x_j,\dots,x_{n-2})},$$

and

$$\frac{\partial C_{x_{n-1},x_j|x_1,\dots,x_j,\dots,x_{n-2}}(F(x_{n-1}|x_1,\dots,x_j,\dots,x_{n-2}),F(x_j|x_1,\dots,x_j,\dots,x_{n-2}))}{\partial F(x_j|x_1,\dots,x_j,\dots,x_{n-2})}.$$

To minimize the number of distributions to compute we choose the same index j in the two previous formulas. Thus, we have three new terms to evaluate:

$$F(x_k|x_1,\dots,x_j,\dots,x_{n-2}),\ k=n,n-1,j. \tag{3}$$

- We need now to iterate the formula (2) to know which terms are required at the next step. We will use four indices: α_1,β_1 for $f(x_1,\dots,x_{n-1})$ and α_2,β_2 for $f(x_1,\dots,x_{n-2},x_n)$. Thus, we compute:

$$F(x_{\alpha_i}|x_1,\dots,x_{\alpha_i},\dots,x_{\beta_i},\dots,x_{n-2}),i=1,2 \tag{4}$$

and

$$F(x_{\beta_i}|x_1,\dots,x_{\alpha_i},\dots,x_{\beta_i},\dots,x_{n-2}),i=1,2. \tag{5}$$

By matching the three previous formulas (3) and the four new ones (4)-(5), we chose $\alpha_1=n-1,\beta_1=j,\alpha_2=n,\beta_2=j$, so as to minimize the number of new distributions to compute.

Unwinding the rest of the algorithm with the same choice of indices at each step leads to regular vines, insuring us that the regular vines are those with the less pair-copulas. This characteristic is very interesting but it does not permit to say that the regular vines are the best copulas.

4.3 Convergence

Let $X=(X_i)$ for $i=1,\cdots n$ such that for all i we have a T sample. Let C_{θ_0} be the copulas associated with the vector X. We assume that the dimension of θ_0 is k and it is constant for all vines. We assume that $\forall i\ X_{i,1}\dots X_{i,T}$ are independent identically distributed random variables[5]. To estimate θ_0 we use a n-stage maximum likelihood procedure yielding $\hat{\theta}_T$. We state that for all the

[5]To achieve this one can mount univariate models on each X_i.

vine formulas we have the same rate of convergence $\sqrt{T}$ and that there is no diagonal dominance of any vines, thus:

$$M \cdot \sqrt{T}\left(\hat{\theta}_T - \theta_0\right) \xrightarrow{D} N(0, I_k),$$

where M is a bounded $n \times n$ matrix, (Guégan and Maugis, 2008). If it happens that the same copula is used multiple times in the estimation, then the parameters of $\hat{\theta}_T$ are set accordingly: this does not affect the previous convergence. We compare the estimation of each distribution function $F(.|.)$ instead of the whole model. Indeed, these are estimated with the same number of parameters and the same number of maximum-likelihood stages. The reduction of parameters only appears as the estimation of such terms overlap.

This means that, whatever the number of parameters needed to estimate a vine formula, it will have the same convergence speed. This means that since we have a bigger choice of vines we can potentially describe more varied behaviors.

5 Model Selection

In this section we show an efficient search algorithm to select one estimator among all the vines we have considered previously. To do so we use lattice, which is a partial order on the set of vines formulas. The idea is the following: among models, some can be considered as specifications of others, and some as generalizations of others. For instance, consider the following example. Let X, Y and Z be three random variables: the linear regression model $Z = \alpha X$ is more general than the linear regression model $Z = \alpha' X + \beta Y$. Conversely, the representation $Z = \alpha' X + \beta Y$ is a specification of the model $Z = \alpha X$. In the following we will use the term specification in this sense.

5.1 Lattice Selection

A lattice is built on a binary relationship. For instance, the real numbers are a lattice with respect to the order relation $x < y$ and set of subsets is a lattice for the inclusion relation $X \subset Y$. Lattice are more complex structures, but this last approach is sufficient for our purpose.

The use of lattices to organize large sets of models is not new but lattice selection has never been used on vines. Gabriel (1969) set down the principles and gave the theoretical ground for such method. Practically we assume that specifications of a false model are false, and that generalizations of valid models are valid. A general algorithm can be found in Edwards and Havranek (1987).

The method relies on a decision rule which permits to decide whether a model if acceptable or not. For this purpose many test exists. Here, we retain the

"alternative test" described in Chen et al. (2004) because it focuses on the dependence structure of the data set and not uniquely on the likelihood or on the fit.

5.2 Lattices on Vine

To create specifications and sub-models we can allow for each pair-copula within a vine to be either the estimated copula or the independent copula, then the more the copulas are set to be independent the simpler the model is. This approach would be the most efficient lattice structure for vine model search, however it is mostly not feasible as yet because it generates too many different models. Thus, we need to consider a simpler lattice.

Vines are the product of many copulas: first the copulas with no constraint $c_{1,2}(F(x_1), F(x_2))$, then those with one constraint $c_{1,2|3}(F(x_1|x_3), F(x_2|x_3))$, until those with $(n-2)$ constraints $c_{1,2|3\ldots n}(F(x_1|x_3\ldots x_n), F(x_2|x_3\ldots x_n))$. Using this remark, we can build the following lattice. First, for each vine formula we decide that all copulas with one or more than one constraint are the independent copulas. Then, all the copulas with more than two constraints and up are the independent copulas and so on. Thus, the more copulas are set to be independent, the simplest the model is.

Other lattice structures are possible, we can use Gaussian copula or partial independence, in that case the lattice structure will have different search power and will select different models. It would be important to obtain an optimum lattice through an efficient search.

6 Conclusion

In this note we have described various ways to use vines, allowing the user to make best use of its capacities. One can select the set of vine formulas and of bi-variate copulas for the estimation, and choose the lattice structure and the decision rule for the selection. This methodology leaves a lot of degrees of freedom for the user and a lot of questions. What is the best criterion to decide which option to take in which cases? Should one use many different pair-copulas if the data are not very correlated? Are un-regular vines necessary if the dependence is simple? Which decision rule with which lattice structure? The complexity of these questions are twofold.

First, the complexity of the possible vines models and the number of vines makes a complete examination of them and the behavior they describe nearly impossible, so that the analysis require new theoretical tool to compare them. Second, devising this tool is complex because of the ambiguous status of vines:

between estimators and models of their own so that usual comparison tools (speed of convergence, likelihood, ...) are not efficient.

The computational wall is high: to compute the tails parameters of all the vines one needs n^2 quotient of $(n-1)$-th degree integrals for each $\prod_{i=1}^{n} i!$ possible vines. Since we are concerned with cases such as $n \approx 100$, $1s$ per integral would mean more than 10^{300} years of computation!

Vine copulas stretch the capacity of classic pair-copulas, and all the issues from estimation to selection, the resulting computational problems and possible over-estimation are consequences of us trying to compensate for this lack. One could feel that research of more valid and/or more varied pair-copulas is necessary: it is the case, but it has already gone on for a long time. What is important is that the best use of vines is still to be found, it has clearly high potential, Aas et al. (2007) present a reasonable use, we presented many new ways it could be used and there exist surely many more. There remains to develop proper evaluation and comparison tools to devise an optimum way to use vines. All this makes vines a very interesting and promising subject even if it is a arduous one.

References

K. Aas, C. Czadob, Frigessic, and H. A., Bakkend. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 2007.

T. Bedford and R.M. Cooke. Vines: A New Graphical Model for Dependent Random Variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

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

D. Berg and K. Aas. Models for construction of multivariate dependence. *Working Paper*, 2007.

X. Chen, Y. Fan, and A.J. Patton. Simple Tests for Models of Dependence Between Multiple Financial Time Series, with Applications to U.S. Equity Returns and Exchange Rates. *Working Paper*, 2004.

D. Edwards and T. Havranek. A Fast Model Selection Procedure for Large Families of Models. *Journal of the American Statistical Association*, 82(397): 205–213, 1987.

K. R. Gabriel. Simultaneous Test Procedures - Some Theory of Multiple Comparisons. *Annals of Mathematics Statistics*, 40(1):224–250, 1969.

D. Guégan and P. A. Maugis. Dealing With Vines Issues. *Working Paper*, 2008.

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, 1997.