A vine and gluing copula model for permeability stochastic simulation

Arturo Erdely[1] and Martin Diaz-Viera[2]

[1] Actuarial Science Program, Facultad de Estudios Superiores Acatlan, Universidad Nacional Autonoma de Mexico, Mexico
(E-mail: arturo.erdely@comunidad.unam.mx)

[2] Gerencia de Ingenieria de Recuperacion Adicional, Instituto Mexicano del Petroleo, Mexico
(E-mail: mdiazv@imp.mx)

Abstract. Statistical dependence between petrophysical properties in heterogeneous formations is usually nonlinear and complex; therefore, traditional statistical methods based on assumptions of linearity and normality are usually not appropriate. Copula based models have been previously applied to this kind of variables but it seems to be very restrictive to find a single copula family to be flexible enough to model complex dependencies in highly heterogeneous porous media. The present work combines vine copula modeling with a bivariate gluing copula approach to model rock permeability using vugular porosity and measured P-wave velocity as covariates in a carbonate double-porosity formation at well log scale.
Keywords: Vine and gluing copulas, nonlinear dependence, petrophysical modeling.

1 Copula basics

A *copula function* is the functional link between the joint probability distribution function of a random vector and the marginal distribution functions of the random variables involved. For example, in a bivariate case, if (X,Y) is a random vector with joint probability distribution $F_{XY}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$ with continuous marginal distribution functions F_X and F_Y then by *Sklar's Theorem*[19] there exists a unique bivariate copula function $C_{XY} : [0,1]^2 \to [0,1]$ such that $F_{XY}(x,y) = C_{XY}(F_X(x), F_Y(y))$. Therefore, all the information about the dependence between X and Y is contained in the underlying copula C_{XY}, since F_X and F_Y only explain the individual (marginal) behavior of such random variables. As an example, for continuous random variables, X and Y are independent if and only if $C_{XY}(u,v) = \Pi(u,v) := uv$.

As a consequence of results by Hoeffding[9] and Fréchet[6], particularly what is known as the *Fréchet-Hoeffding bounds* for joint probability distribution functions, Sklar's Theorem leads to the following sharp bounds for any bivariate copula: $W(u,v) \leq C_{XY}(u,v) \leq M(u,v)$ for all u,v in $[0,1]$, where $W(u,v) := \max\{u+v-1, 0\}$ and $M(u,v) := \min\{u,v\}$ are themselves copulas. W (respectively M) is the underlying copula of a bivariate random vector of

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

continuous random variables (X, Y) if, say, Y is an almost surely decreasing (respectively increasing) function of X.

Formal definitions and main properties of copula functions are covered in detail in Nelsen[14] and Durante and Sempi[3]. Among many other properties, any copula C is a uniformly continuous function, and in particular its *diagonal section* $\delta_C(t) := C(t,t)$ is uniformly continuous and nondecreasing on $[0,1]$. In terms of the Fréchet-Hoeffding bounds, we get $\max\{2t-1, 0\} \leq \delta_C(t) \leq t$ for all t in $[0,1]$.

Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote an observed sample of size n from a bivariate random vector (X, Y) of continuous random variables. We may estimate the underlying copula C_{XY} by the *empirical copula* C_n, see Deheuvels[2], which is a function with domain $\{\frac{i}{n} : i = 0, 1, \ldots, n\}^2$ defined as:

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) := \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{rank(x_k) \leq i\,,\, rank(y_k) \leq j\} \tag{1}$$

and its convergence to the true copula C_{XY} has also been proved, see Rüschendorf[17] and Fermanian *et al.*[5]. Strictly speaking, the empirical copula is not a copula since it is only defined on a finite grid, but by Sklar's Theorem C_n may be extended to a copula. Based on the empirical copula several goodness-of-fit tests have been developed, see for example Genest *et al.*[7], to choose the best parametric family of copulas from an already existing long catalog, see for example chapter 4 in Joe[11].

The underlying copula C_{XY} is invariant under strictly increasing transformations of X and Y, that is $C_{XY} = C_{\alpha(X),\beta(Y)}$ for any strictly increasing functions α and β. Recall that for any continuous random variable X we have that the random variable $F_X(X)$ is uniformly distributed on the open interval $]0,1[$. Let $U := F_X(X)$ and $V := F_Y(Y)$, then (X, Y) has the same underlying copula as (U, V) and by Sklar's Theorem $F_{UV}(u,v) = C_{UV}(F_U(u), F_V(v)) = C_{UV}(u,v)$. So the transformed sample $\{(u_1, v_1), \ldots, (u_n, v_n)\}$ where $(u_k, v_k) = (F_X(x_k), F_Y(y_k))$ may be considered as *observations* from the underlying copula C_{XY}. If F_X and F_Y are unknown (which is usually the case) they can be replaced by the empirical aproximation $F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{x_k \leq x\}$ and in such case we obtain what is known as *pseudo-observations* of the underlying copula C_{XY}, which are used for copula estimation purposes, since they are equivalent to the ranks in (1).

2 Gluing copulas

Sklar's Theorem is also useful for building new multivariate probability models. For example, if F and G are univariate probability distribution functions, and C is any bivariate copula, then $H(x,y) := C(F(x), G(y))$ defines a joint probability distribution function with univariate marginal distributions F and G. Several methods for constructing families of copulas have been developed (geometric methods, archimedean generators, ordinal sums, convex sums, shuffles) and among them we may include *gluing copulas* by Siburg and Stoimenov[18], which we will illustrate in a very particular case: let C_1 and C_2 be two given bivariate copulas, and $0 < \theta < 1$ a fixed value, we may scale C_1 to $[0, \theta] \times [0, 1]$ and C_2 to $[\theta, 1] \times [0, 1]$ and *glue* them into a single copula:

$$C_{1,2,\theta}(u, v) := \begin{cases} \theta C_1(\frac{u}{\theta}, v), & 0 \leq u \leq \theta, \\ (1-\theta)C_2(\frac{u-\theta}{1-\theta}, v) + \theta v, & \theta \leq u \leq 1. \end{cases} \tag{2}$$

A gluing copula construction may easily lead to a copula with a diagonal section $\delta_{1,2,\theta}(t) = C_{1,2,\theta}(t, t)$ that has a discontinuiy in its derivative at the *gluing point* $t = \theta$. This fact may be taken into consideration when trying to fit a parametric copula to observed data, since common families of copulas have diagonal sections without discontinuities in their derivatives, and if the *empirical diagonal* $\delta_n(\frac{i}{n}) := C_n(\frac{i}{n}, \frac{i}{n})$ strongly suggests there is one or more points at which a discontinuity of the derivative occurs, an appropriate data partition by means of finding some gluing points could be helpful to model the underlying copula by the gluing copula technique.

For a more specific example, in the particular case $C_1 = M$ and $C_2 = \Pi$ it is straightforward to verify that for $0 \leq t \leq \theta$ we get a diagonal section $\delta_{1,2,\theta}(t) = \theta t$, while for $\theta \leq t \leq 1$ we get $\delta_{1,2,\theta}(t) = t^2$ and clearly the left and right derivatives at the gluing point $t = \theta$ are not the same, see Figure 1.

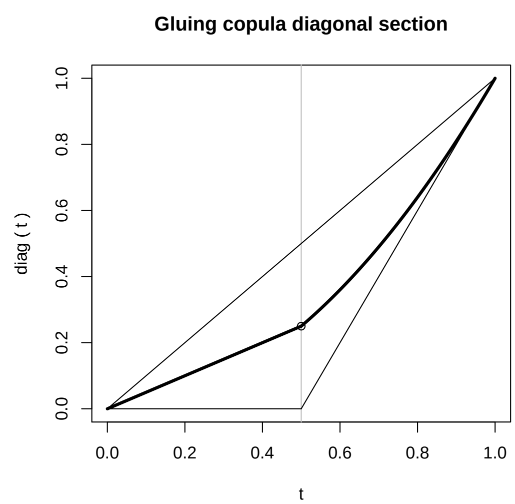

Fig. 1. Diagonal section of the resulting gluing copula with $C_1 = M$, $C_2 = \Pi$ and gluing point $\theta = \frac{1}{2}$

3 Trivariate vine copulas

In the previous sections we summarized some main facts about bivariate copulas, but Sklar's Theorem is valid for any $d \geq 2$ random variables. For example, in the case of a trivariate random vector (X_1, X_2, X_3) of continuous random variables with joint probability distribution $F_{123}(x_1, x_2, x_3) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, X_3 \leq x_3)$ and marginal univariate distributions F_1, F_2, and F_3, by Sklar's Therem there exists a unique underlying copula $C_{123} : [0,1]^3 \to [0,1]$ such that $F_{123}(x_1, x_2, x_3) = C_{123}(F_1(x_1), F_2(x_2), F_3(x_3))$. In case F_{123} is *absolutely continuous* we may obtain the following expression for the trivariate joint density:

$$f_{123}(x_1, x_2, x_3) = c_{123}(F_1(x_1), F_2(x_2), F_3(x_3))f_1(x_1)f_2(x_2)f_3(x_3) \tag{3}$$

where the copula density $c_{123}(u, v, w) = \frac{\partial^3}{\partial u \partial v \partial w} C_{123}(u, v, w)$ and the marginal densities $f_k(x) = \frac{d}{dx} F_k(x)$, $k \in \{1, 2, 3\}$. According to Kurowicka[13]:

> The choice of copula is an important question as this can affect the results significantly. In the bivariate case $[d = 2]$, this choice is based on statistical tests when joint data are available [...] Bivariate copulae are well studied, understood and applied [...] Multivariate copulae $[d \geq 3]$ are often limited in the range of dependence structures that they can handle [...] Graphical models with bivariate copulae as building blocks have recently become the tool of choice in dependence modeling.

The main idea behind *vine copulas* (or pair-copula constructions) is to express aribitrary dimensional dependence structures in terms of bivariate copulas and univariate marginals. For example, we may rewrite the trivariate joint density (3) in the following manner by conditioning in one of the random variables, say X_1 :

$$\begin{aligned} f_{123} &= f_{23|1} \cdot f_1 \\ &= c_{23|1}(F_{2|1}, F_{3|1}) \cdot f_{2|1} \cdot f_{3|1} \cdot f_1 \\ &= c_{23|1}(F_{2|1}, F_{3|1}) \cdot \frac{f_{12}}{f_1} \cdot \frac{f_{13}}{f_1} \cdot f_1 \\ &= c_{23|1}(F_{2|1}, F_{3|1}) \cdot c_{12}(F_1, F_2) \cdot c_{13}(F_1, F_3) \cdot f_1 \cdot f_2 \cdot f_3 \end{aligned} \tag{4}$$

with other two similar possibilities by conditioning on random variables X_2 or X_3. If $\{(x_{1k}, x_{2k}, x_{3k})\}_{k=1}^n$ is a an observed sample size n from an absolutely continuous random vector (X_1, X_2, X_3) we may use the bivariate observations $\{(x_{1k}, x_{2k}\}_{k=1}^n$ to estimate c_{12} and $F_{2|1}$, and we use $\{(x_{1k}, x_{3k}\}_{k=1}^n$ to estimate c_{13} and $F_{3|1}$. Following the ideas in Gijbels *et al.*[8] we obtain the following expression for the conditional bivariate joint distribution of (X_2, X_3) given $X_1 = x_1$:

$$\begin{aligned} F_{23|1}(x_2, x_3 \,|\, x_1) &= \mathbb{P}(X_2 \leq x_2, X_3 \leq x_3 \,|\, X_1 = x_1) \\ &= C_{23|1}\big(F_{2|1}(x_2 \,|\, x_1), F_{3|1}(x_3 \,|\, x_1) \,\big|\, x_1\big) \end{aligned} \tag{5}$$

Here the value x_1 becomes a parameter for the conditional bivariate copula $C_{23|1}$ and for the conditional univariate marginals $F_{2|1}$ and $F_{3|1}$. In case there is some kind of evidence (empirical or expert-based) to assume that the underlying bivariate copula for $F_{23|1}$ does not depend on the value of the conditioning variable, we have what is known as a *simplifying assumption*, see for example Joe[11], and so to estimate such bivariate copula $C_{23}^{*} \equiv C_{23|1}$ again we may follow the ideas in Gijbels *et al.*[8] and use the pseudo-observations $\{(u_{2k}, u_{3k}) = \big(F_{2|1}(x_{2k} \,|\, x_{1k}), F_{3|1}(x_{3k} \,|\, x_{1k})\big)\}_{k=1}^{n}$.

4 Application to petrophysical data

As mentioned in Erdely and Diaz-Viera[4]:

> Assessment of rock formation permeability is a complex and challenging problem that plays a key role in oil reservoir modeling, production forecast, and the optimal exploitation management [...] Dependence relationships [among] petrophysical random variables [...] are usually nonlinear and complex, and therefore those statistical tools that rely on assumptions of linearity and/or normality and/or existence of moments are commonly not suitable in this case.

In the present work we apply a trivariate vine copula model to petrophysical data from Kazatchenko *et al.*[12] for variables X_1 = vugular porosity (PHIV), X_2 = measured P-wave velocity (VP), and X_3 = permeability(K), see Figure 2 for bivariate scatterplots and bivariate copula pseudo-observations.

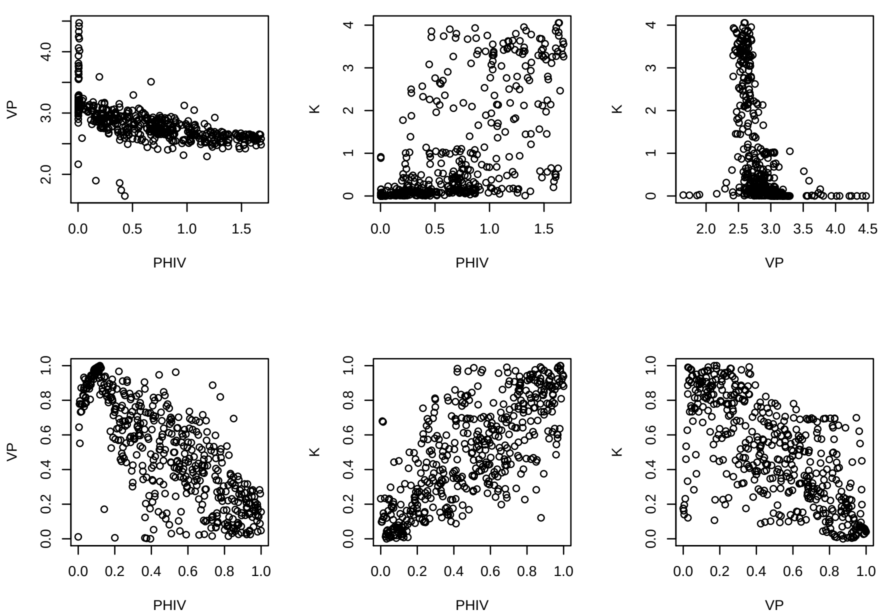

Fig. 2. *First row:* bivariate scatterplots. *Second row:* bivariate copula pseudo-observations.

First we searched for empirical evidence to check if a simplifying assumption is reasonable by splitting the pseudo-observations

$\{(u_{2k}, u_{3k}) = \big(F_{2|1}(x_{2k} \,|\, x_{1k}), F_{3|1}(x_{3k} \,|\, x_{1k})\big)\}_{k=1}^{n}$ in two sets A and B depending on whether the conditioning variable was less or greater than its median, and use them for an equality of copulas hypothesis test $\mathcal{H}_0 : C_A = C_B$ by Rémillard and Scaillet[16] implemented in the `TwoCop` R-package[15], see Table 1 for a summary of the results obtained. An extremely low p-value leads to the conclusion of rejecting a simplifying assumption, since lower values of the conditioning variable suggest a different dependence structure that the one corresponding to higher values. From Table 1 we conclude that a simplfying assumption condtioning on variable X_3 is definitely rejected, and conditioning on X_1 would be the best option in this case.

Conditioning variable	Simplifying assumption p-value
X_1	0.34
X_2	0.13
X_3	0.00

Table 1. p-values from Rémillard-Scaillet test adapted to test for simplifying assumption.

For the three bivariate copulas needed in the trivariate vine copula model (4) no single family of parametric bivariate copulas was able to achieve an acceptable goodness-of-fit, according to results obtained with the `copula` R-package[10]. Therefore a gluing approach has been applied, using a heuristic procedure to find gluing point candidates, called also *knots*, for a piecewise cubic polynomial fit (a particular case of *splines*) to the empirical diagonal δ_n but without the usual assumption of having continuous first or second derivative at the knots, since for gluing copula purposes that is exactly what we are looking for: points of discontinuity in the derivative of the diagonal section of the underlying copula.

Let $\mathcal{K} := \{t_0, \ldots, t_m\}$ be a set of $m + 1$ knots in the interval $[0, 1]$ such that $0 = t_0 < t_1 < \cdots < t_m = 1$. Consider the set $\mathcal{P}$ of all continuous functions p on $[0, 1]$ such that:

- $p(t_i) = \delta_n(t_i)$, $i \in \{0, 1, \ldots, m\}$
- p is a cubic polynomial on $[t_{i-1}, t_i]$, $i \in \{1, \ldots, m\}$

The goal is to find the smallest sets of knots $\mathcal{K}$ such that the *mean squared error (MSE)* of piecewise polynomial approximations to each empirical diagonal δ_n is minimal and such that it is possible to reach an acceptable goodness-of-fit of bivariate copulas for the data partitions induced by each $\mathcal{K}$:

Step 1 Calculate pseudo-observations $\mathcal{S} := \{(u_k, v_k) : k = 1, \ldots, n\}$ and rearrange pairs such that $u_1 < \cdots < u_n$.

Step 2 Calculate empirical diagonal $\mathcal{D}_n := \{(\frac{i}{n}, \delta_n(\frac{i}{n})) : i = 0, 1, \ldots, n\}$.

Step 3 Find optimal knot (or gluing point) $t^* = \frac{i^*}{n}$ such that $\mathcal{K} = \{0, t^*, 1\}$ leads to minimal MSE on $\mathcal{D}_n$.

Step 4 Define subsets $\mathcal{G}_1$ and $\mathcal{G}_2$ from $\mathcal{S}$ such that $\mathcal{G}_1 := \{(u_k, v_k) \in \mathcal{S} : u_k \leq t^*\}$ and $\mathcal{G}_2 := \{(u_k, v_k) \in \mathcal{S} : u_k \geq t^*\}$.

Step 5 Apply goodnes-of-fit tests for parametric copulas in each subset $\mathcal{G}_1$ and $\mathcal{G}_2$.

Step 6 If an acceptable fit is reached in both subsets, we are done. Otherwise, apply steps 1 to 5 to the subset(s) which could not fit.

In Table 2 we present a summary of results, specifying how many partitions were needed and the best copula goodness-of-fit achieved on each one, for each bivariate relationship required by (4), making use of the `copula` R-package[10].

Bivariate dependence	Best parametric copula fit	p-value
$X_1\,,\,-X_2$	Plackett*	0.6079
	Galambos*	0.1384
	Plackett	0.3941
	independence	0.5200
$X_1\,,\,X_3$	Plackett*	0.6539
	Clayton	0.1494
	Husler-Reiss	0.8586
$-X_2\,,\,X_3 \,\vert\, X_1$	Plackett*	0.3541
	Clayton*	0.4800

Table 2. Families of copulas indicated with * means that the transformed copula $C^*(u,v) = u + v - 1 + C(1-u, 1-v)$ was used, where C is the original copula family.

5 Final remark

According to Czado and Stöber[1]:

> [...] compared to to the scarceness of work on multivariate copulas, there is an extensive literature on bivariate copulas and their properties. Pair copula constructions (PCCs) build high-dimensional copulas out of bivariate ones, thus exploiting the richness of the class of bivariate copulas and providing a flexible and convenient way to extend the bivariate theory to arbitrary dimensions.

But even expecting a single copula family to be able to model a complex bivariate dependency seems to be still too restrictive, at least for the petrophysical variables under consideration in this work. In such case, an alternative found was to apply a gluing copula approach[18]: decomposing bivariate samples into subsamples whose dependence structures were simpler to model by known parametric families of copulas, taking advantage of already existing tools (and their computational implementations) for bivariate copula estimation.

Acknowledgement

The present work was supported by project IN115914 from Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) at Universidad Nacional Autónoma de México.

References

1. C. Czado and J. Stöber. Pair Copula Constructions. In J.-F. Mai and M. Scherer, editors, em Simulating Copulas, pp. 185–230, Imperial College Press, London, 2012.
2. P. Deheuvels. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci.*, 65, 5, 274–292, 1979.
3. F. Durante and C. Sempi. *Principles of Copula Theory*, CRC Press, Boca Raton, 2016.
4. A. Erdely and M. Diaz-Viera. Nonparametric and Semiparamegric Bivariate Modeling of Petrophysical Porosity-Permeability Dependence from Well Log Data. In P. Jaworski, F. Durante, W. Härdle, T. Rychlik, editors, *Copula Theory and Its Applications*, pp. 267–278, Springer-Verlag, Berlin Heidelberg, 2010.
5. J-D. Fermanian, D. Radulovíc, M. Wegcamp. Weak convergence of empirical copula processes. *Bernoulli*, 10, 847–860, 2004.
6. M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon*, 14, (Sect. A Ser.3), 53–77, 1951.
7. C. Genest, B. Rémillard, D. Beaudoin. Goodness-of-fit tests for copulas: a review and a power study. *Insurance Math. Econom.*, 44, 199–213, 2009.
8. I. Gijbels, N. Veraverbeke, M. Omelka. Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis,* 55, 1919–1932, 2011.
9. W. Hoeffding. Masstabinvariante Korrelationstheorie. *Schriften des Matematischen Instituts und des Instituts für Angewandte Matematik der Universität Berlin*, 5, 179–223, 1940.
10. M. Hofert, I. Kojadinovic, M. Maechler, J. Yan. copula: Multivariate Dependence with Copulas. R package version 0.999-13, URL http://CRAN.R-project.org/package=copula , 2015.
11. H. Joe. *Dependence Modeling with Copulas*, CRC Press, Boca Raton, 2015.
12. E. Kazatchenko, M. Markov, A. Mousatov, J. Parra. Carbonate microstructure determination by inversion of acoustic and electrical data: application to a South Florida Aquifer. *J. Appl. Geophys.*, 59, 1–15, 2006.
13. D. Kurowicka. Introduction: Dependence Modeling. In D. Kurowicka and H. Joe, editors, *Dependence Modeling Vine Copula Handbook*, pp. 1–17, World Scientific Publishing, Singapore, 2011.
14. R. B. Nelsen. *An introduction to copulas*, Springer, New York, 2006.
15. B. Rémillard and J.-F. Plante. TwoCop: Nonparametric test of equality between two copulas. R package version 1.0. http://CRAN.R-project.org/package=TwoCop , 2012.
16. B. Rémillard, B. Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100, 377–386, 2009.
17. L. Rüschendorf. Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.*, 4, 912–023, 1976.

18. K. F. Siburg and P.A. Stoimenov. Gluing copulas. *Commun. Statist.– Theory and Methods*, 37, 3124–3134, 2008.
19. A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8, 229–231, 1959.