Modeling Dependence with Copulas: Theory and Applications, Quo Vadis?

Hui Lin

December 3, 2012

Contents

1 Introduction

The word "copula" as a grammatical term for a word or expression that links a subject and predicate, Sklar felt that this would make an appropriate name for a function that links a multidimensional distribution to its one-dimensional margins, and used it as such. The history of copulas may be said to begin with Frechet (1951). Frechet's problem: given the distribution fuctions F_j with $j = 1, ..., d$ of d r.v's $X_1, X_2, ..., X_d$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, what can be said about the set $\Gamma(F_1, F_2, ..., F_d)$ of the $d-$dimensional distribution functions whose marginals are the given F_j?

$$H \in \Gamma(F_1, ..., F_d) \Leftrightarrow H(+\infty, ..., +\infty, t, +\infty, ..., +\infty) = F_j(t) \tag{1}$$

The set $\Gamma(F_1, ..., F_d)$ is called the Frechet class of the $F_j's$. Notive $\Gamma(F_1, ..., F_d) \neq \varnothing$ since, if $X_1, X_2, ..., X_d$ are independent, then

$$H(x_1, x_2, ..., x_d) = \prod_{j=1}^{d} F_j(x_j)$$

But, it was not clear which the other elements of $\Gamma(F_1, ..., F_d)$ were. Dall' Aglio (1972) studied the conditions under which there is just one distribution function belonging to $\Gamma(F_1, F_2)$. At the end of the nineties, the notion of copulas became increasingly popular due to an explosive development of quantitative risk management methodology within finance and insurance. Two papers more than any others put the fire to the fuse: Embrechts et al.(2002) and Li credit portfolio model(Li2001).

When a r.v $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ is given, two problems are interesting:

1. the probabilistic behaviour of each one of its components

2. the relationship among them

The main goal of the study is to show how copulas allow to answer the second question. The outline of this report is as follows. In section 2 we introduced some definitons and classical copula models where we performed several simulation studies and presented the densities of some copula in graphic. Then we talked about some possible ways to estimate copula in section 3. For bivariate case, we illustrated a parametric method using piesewise linear

approximation which is explained in section 4. At last, in section 5, we applied the methods introduced in previous sections to dietary intake and compared the results.

2 Copula Definitions and Classical Copula models

In this section, we give some fundamental definitions regarding copula and introduce some important classes of copulas.

2.1 Basic definitions and examples

Definition 2.1. A d-dimensional copula $C : [0,1]^d \longrightarrow [0,1]$ is a function which is a cumulative distribution function with uniform marginals.

Theorem 2.2. ***(Sklar1959)*** *Consider a d-dimensional cdf H with marginals $F_1, ..., F_d$. There exists a copula C, such that*

$$H(x_1, ..., x_p) = C(F_1(x_1), ..., F_d(x_d)) \tag{2}$$

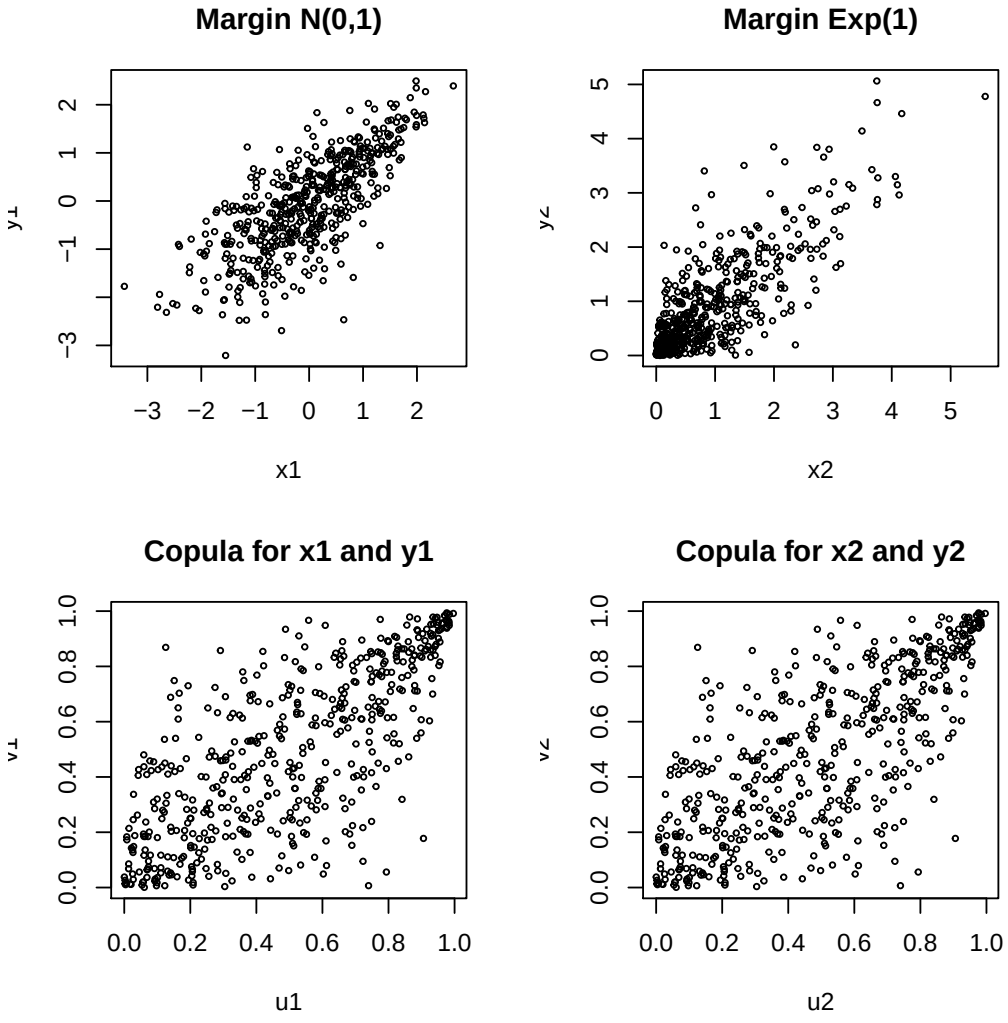
for all $x_i \in \bar{\mathbb{R}}$. If F_i is continuous for all $i = 1, ..., d$ then C is unique; otherwise C is uniquely determined only on $RanF_1 \times \cdots \times RanF_d$, where $RanF_i$ denotes the range of the cdf F_i .

Theorem 2.3. ***(Rank invariant)*** *Let $X = (X_1, ..., X_d)$ be a r.v. with continuous H , univariate marginals $F_1, F_2, ..., F_d$, and copula C. Let $T_1, ..., T_d$ be strictly increasing functions from$\mathbb{R}$ to$\mathbb{R}$. Then C is also the copula of $(T_1(X_1), T_2(X_2), ..., T_d(X_d))$.*

A direct consequence of Theorem 2.3 is that copula properties are invariant under strictly increasing transformations of the underlying random variables. This seems, at first sight, to be counterintuitive because monotone transformations of course will change the dependence. But after removing the effect of the marginals we end up with the same dependence structure. Consider the following example.

```
> rgumbelCopula(500,alpha=2)->rgum
> qnorm(rgum[,1])->norm1
> qnorm(rgum[,2])->norm2
```

```
> qexp(rgum[,1])->exp1
> qexp(rgum[,2])->exp2
> par(mfrow=c(2,2),mar=c(4,3,3,2)+0.5)
> plot(norm1,norm2,cex=0.5,xlab="x1",ylab="y1", main="Margin N(0,1)")
> plot(exp1,exp2,cex=0.5,xlab="x2",ylab="y2",main="Margin Exp(1)" )
> plot(rgum,cex=0.5,xlab="u1",ylab="v1",main="Copula for x1 and y1" )
> plot(rgum,cex=0.5,xlab="u2",ylab="v2",main="Copula for x2 and y2")
```



2.2 The Frechet-Hoeffding Bound

Definition 2.4. For any $\mathbf{u} = (u_1, \ldots, u_p)^T \in [0,1]^p$, the functions M_p ,Π_p and W_p defined as follows:

$$M_p(\mathbf{u}) = min_{1 \le i \le p}\{u_i\}$$

$$\Pi_p(\mathbf{u}) = u_1 \times \cdots \times u_p$$

$$W_p(\mathbf{u}) = max\left\{\sum_{i=1}^{p} u_i - p + 1, 0\right\}$$

M_p and Π_p are copulas for any $p \ge 2$. But W_p is only a copula for $p = 2$. Hoeffding and Frechet independently derived that a copula always lies in between certain bounds. The reason for this is the existence of some extreme cases of dependency. Before moving to Frechet-Hoeffding's Bound theorem, let us first consider some extreme cases of dependency to get a little insights.

Example 2.1. Consider U_1 and U_2 are uniform random variables. When $U_1 = U_2$, the two variables are extremely dependent on each other. In this case, the copula for (U_1, U_2) is

$$C(u_1, u_2) = P(U_1 \le u_1, U_2 \le u_2) = min(u_1, u_2) \tag{3}$$

Random variables of this kind are called comonotonic.

Next we consider a contrary example to the above.

Example 2.2. Assume $U_2 = 1 - U_1$, then the related copula is

$$C(u_1, u_2) = P(U_1 \le u_1, 1 - U_1 \le u_2) = P(1 - u_2 \le U_1 \le u_1) = u_1 + u_2 - 1 \tag{4}$$

and 0 otherwise.

Random variables of this kind are called countermonotonic.

Note that copulas in example 2.1 and 2.2 do not have copula densities as they both involve a kink therefore are not differentiable. One has mass only on the diagonal $u_1 = u_2$ and the other on $u_2 = 1 - u_1$. People will naturally consider extend the ideas to multidimentional case. A comonotonic copula exists in any dimension d but there is

no countermonotonic copula when $d \geq 3$. To see this, consider random variables X_1, X_2 *and* X_3. When we set X_1 to be countermonotonic to both X_2 and X_3, this meanwhile restrict relation between X_2 and X_3. More specifically, when X_1 increases, both X_2 and X_3 have to increase, so they can not be countermonotonic again. On the other hand, even a countermonotonic copula does not exist, the bound in the following theorem still holds. See more details, including geometrical interpretations in Mikusinski, Sherwood, and Taylor (1992).

Theorem 2.5. *For any copula* $C(\mathbf{u}) = C(u_1, u_2, ..., u_d)$

$$max\left\{\sum_{i=1}^{d} u_i + 1 - d, 0\right\} \leq C(\mathbf{u} \leq min(u_1, u_2, ..., u_d) \tag{5}$$

2.3 Classes of Copulas

We will in this section discuss copulas derived from some typical multivariate distributions. In this ways, we categorize copulas according to the distributions they are derived from. For example, the multivariate normal distribution will lead to Gaussian copula and the multivariate Student t-distribution leads to the t-copula.

2.3.1 Elliptical Copulas

It is natural to define elliptical copulas as copulas from elliptical distributions. Elliptical distributions share many tractable properties of the multivariate normal distribution. Simulation from elliptical copulas is easy because simulation from elliptical distributions is easy.

Definition 2.6. Gaussian Copulas are copulas with the form

$$C_R^{Ga}(\mathbf{u}) = \Phi_R^n(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_p)) \tag{6}$$

where Φ_R^n denotes the joint distribution function of the $d-$variate standard normal distribution function with linear correlation matrix R and Φ^{-1} denotes the inverse of the distribution function of the univariate standard normal distribution.

When $d = 2$, equation 6 is written as

$$C_R^{Ga}(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left\{-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right\} dsdt \tag{7}$$

where $R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

Assume

$$\mathbf{T} = d\mu + \frac{\mathbf{Z}}{\sqrt{S/\nu}} \tag{8}$$

with $\mu \in \mathbb{R}^d$, $S \sim \chi^2_\nu$ and $\mathbf{Z} \sim N_d(\mathbf{0}, \Sigma)$ are independent, then $\mathbf{T}$ is a $d-$variate $t_\nu-$distribution with mean μ and covariance matrix $\frac{\nu}{\nu-2}\Sigma$ (for $\nu > 2$) . If $\nu \leq 2$, the covariance matrix of $\mathbf{T}$ is not defined.

Definition 2.7. T-copulas are copulas with form

$$C_{\nu,R}^t(\mathbf{u}) = t_{\nu,R}^d(t_\nu^{-1}(u_1), ..., t_\nu^{-1}(u_n)) \tag{9}$$

where $R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$ for $i, j \in \{1, ..., d\}$ and $t_{\nu,R}^d$ is the distribution function of $\frac{\mathbf{Y}}{\sqrt{S/\nu}}$ with $S \sim \chi^2_\nu$ and $\mathbf{Y} \sim N_d(\mathbf{0}, R)$ are independent. Here t_ν is the (equal) margins of $t_{\nu,R}^n$, i.e. the distribution fuction of $\frac{\mathbf{Y}}{\sqrt{S/\nu}}$.

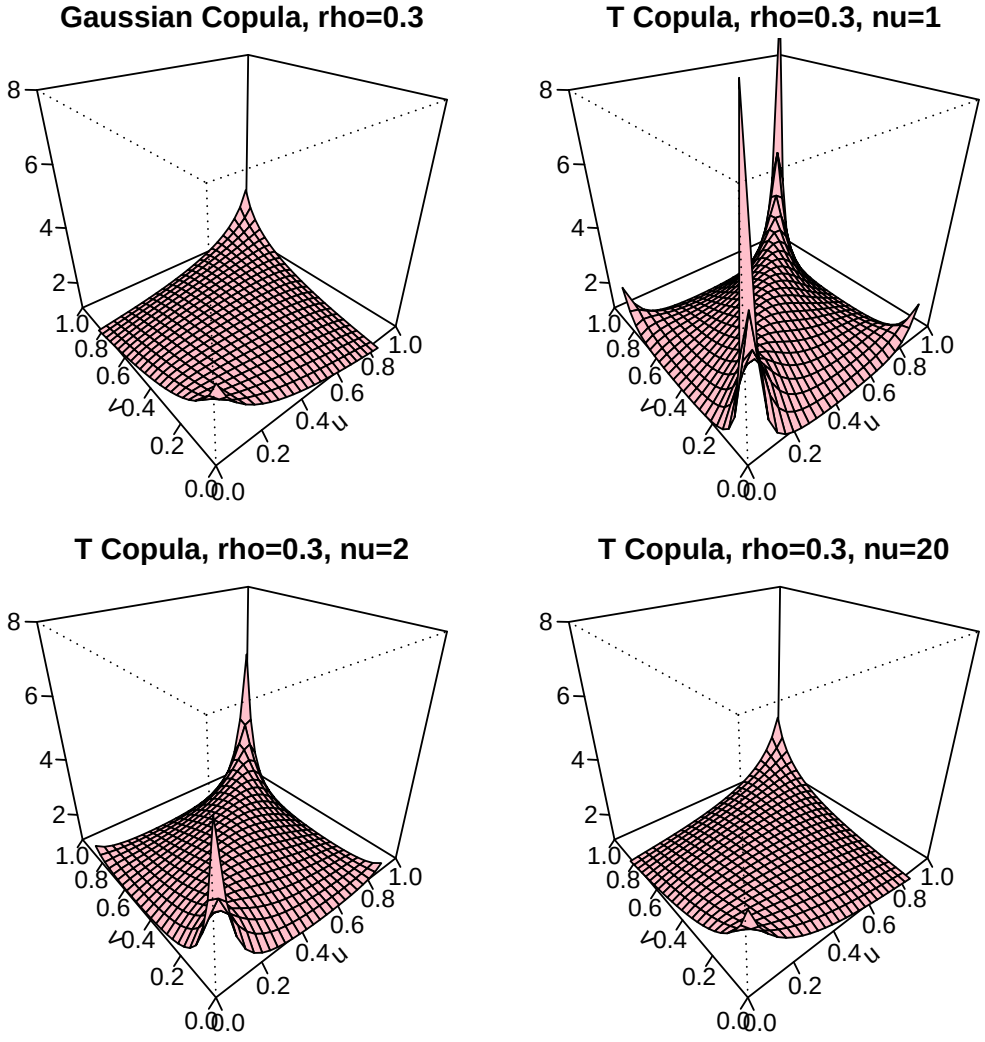
When $d = 2$, equation 9 can be written as

$$C_{\nu,R}^t(u,v) = \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left\{1 + \frac{s^2 2\rho st + t^2}{\nu(1-\rho^2)}\right\}^{-\frac{\nu+2}{2}} dsdt \tag{10}$$

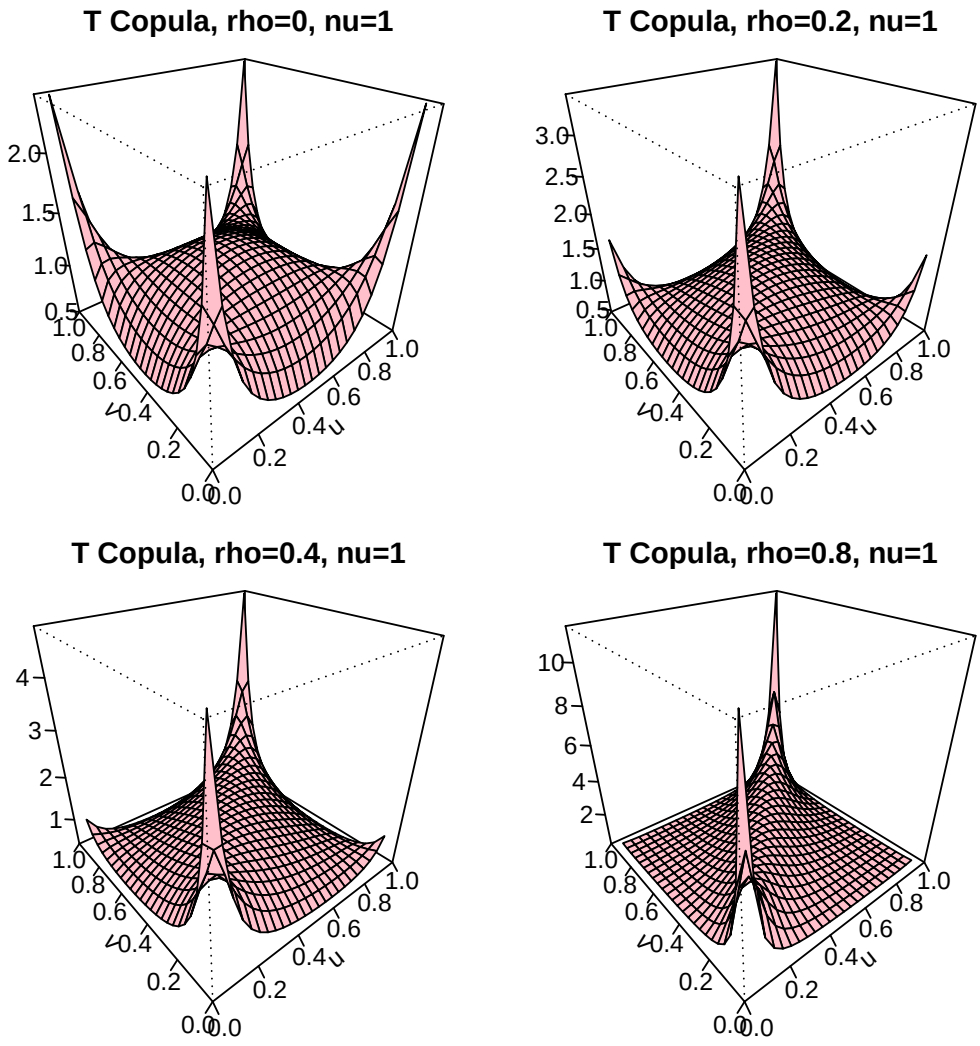
where $R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

The following figure shows the densities of Gaussian copula and a Student t-copula. All have correlation coefficient $\rho = 0.3$. As it can be seen that when ν gets larger, Student t-copula is getting close to Gaussian copula. Also, the behaviour at the four corners is different from the Gaussian copula while they are similar in the center. It indicates that although having the same correlation, the extreme cases (four cornor points) are much more pronounce under t-copula. Consider a finance situation where the random variables describe losses of the protfolio, density at the $(0, 0)$ correspond to big losses in both entities of the portfolio. The fact that t-copula is able to model such

extreme cases is due to the tail-dependence (Def 2.18 2.19 and 2.20).



If we fix ν in t-copula, the following figure shows how the density will change with its mixing nature in bivariate



case. When $\rho = 0$, the density rises up at all four corners symmetrically. Introducing some correlation changes this probability. As the correlation increasing, it is more likely to have values with the same sign which can be easily observed from the peaks at the $(0, 1)$ and $(1, 0)$ corners.

2.3.2 Archimedean Copulas

As mentioned before that since simulation from elliptical distributions is easy, so is simulation from elliptical copulas. However elliptical copulas do not have closed form expressions and are restricted to have radial symmetry($C = \tilde{C}$).

Take bivariate case as example, the density at $(0,0)$ and $(1,1)$ are the same but in many finance and insurance applications it is more reasonable that there is a stronger dependence between big losses (e.g. a stock market crash) than between big gains. Elliptical copulas miss to catch the asymmetries. In this section we will discuss an important class of copulas called Archimedean copulas whcih can be stated directly and have a neat form. Also allowing for a great variety of different dependence structures makes Archimedean copulas more attractive. We wiill just present general definition and a short discussion about Archimedean copulas. More about the topic may be found in Nelsen(1999).

Definition 2.8. Let $\phi : [0,1] \longrightarrow [0,\infty)$ be a strict decreasing function function that satisfies $\phi(0) = \infty$ and $\phi(1) = 0$. Suppose its inverse ϕ^{-1} is completely monotonic on $[0,\infty)$. Then an Archimedean Copula is defined as

$$C(u_1, ..., u_d) = \phi^{-1}\left(\phi(u_1) + ... + \phi(u_d)\right) \tag{11}$$

Definition 2.9. Let the generator function $\phi(u) = \theta^{-1}(u^{-\theta} - 1)$. A Clayton Copula is defined as

$$C_\theta^{Cl}(u_1, u_2, ..., u_d) = \left[\sum_{i=1}^{d} u_i^{-\theta} - d + 1\right]^{-\frac{1}{\theta}}, with\ \theta > 0 \tag{12}$$

Definition 2.10. let the generator function be

$$\phi(u) = -log\left[\frac{exp(-\theta u) - 1}{exp(\theta) - 1}\right] \tag{13}$$

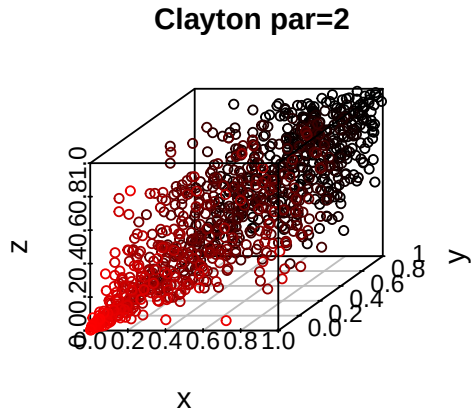
A Frank Copula is defined as

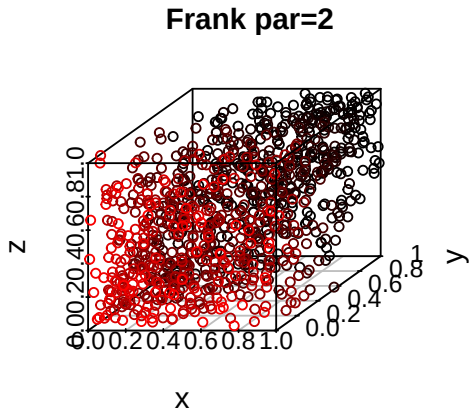
$$C_\theta^{Fr}(u_1, u_2, ..., u_d) = \frac{1}{\theta} log\left\{1 + \frac{\prod_{i=1}^{d}[exp(-\theta u_i) - 1]}{[exp(-\theta) - 1]^{m-1}}\right\} \tag{14}$$

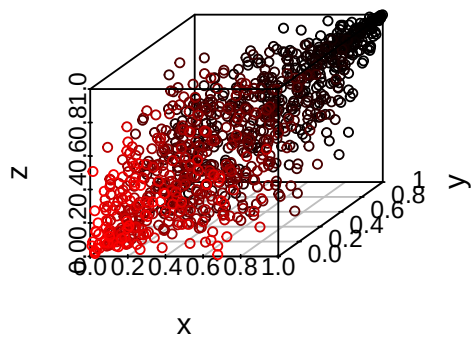
with $\theta \in \mathbb{R}\backslash\{0\}$ for $d = 2$ and $\theta > 0$ for $d \geq 3$.

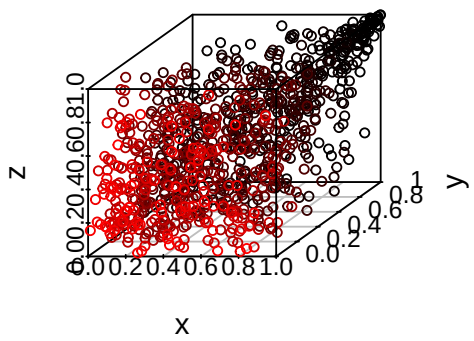
Definition 2.11. Let the generator fuction $\phi(u) = (-logu)^\theta$. A Gumbel Copula is defined as

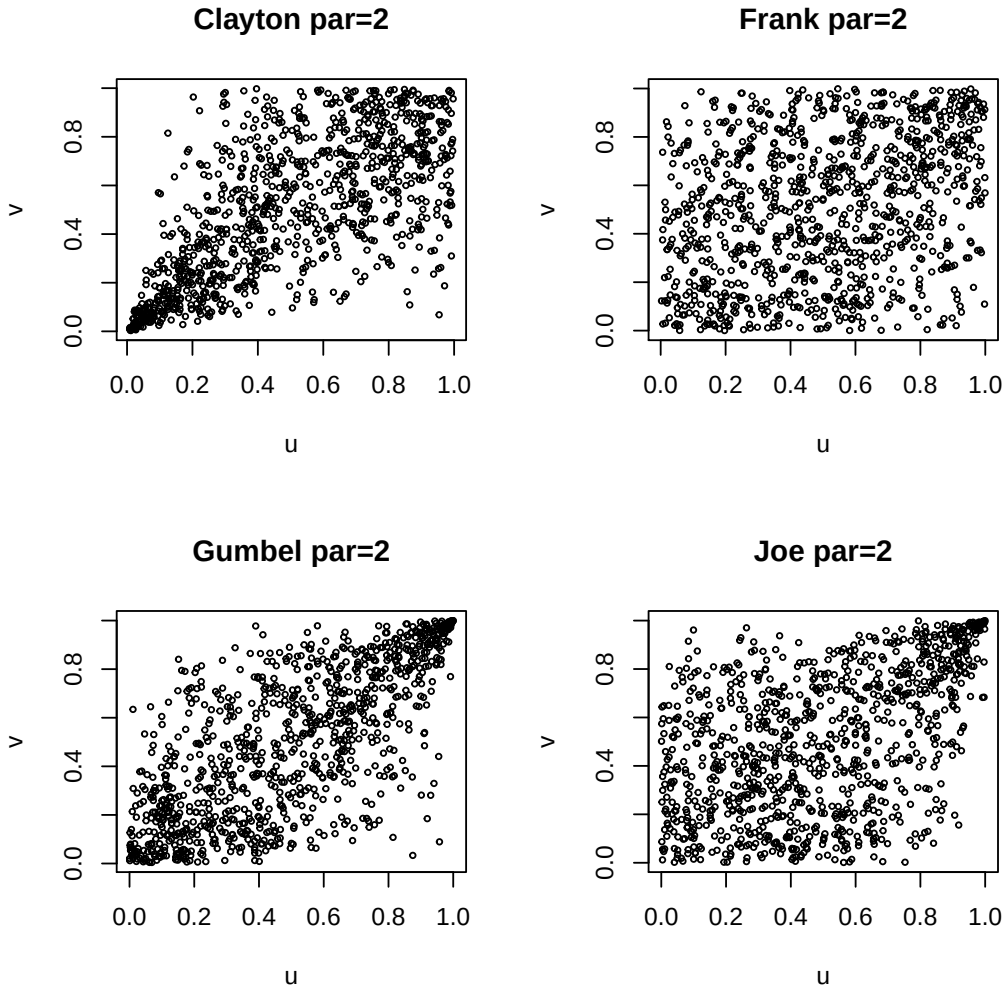
$$C_\theta^{Gu}(u_1, u_2, ..., u_d) = exp\left\{-\left[\sum_{i=1}^{d}(-logu_i)^\theta\right]^{\frac{1}{\theta}}\right\}\ with\ \theta > 1 \tag{15}$$


Clayton par=2
z
x
y
0.0
0.2
0.4
0.6
0.8
1.0


Frank par=2
z
x
y
0.0
0.2
0.4
0.6
0.8
1.0


Gumbel par=2
z
x
y
0.0
0.2
0.4
0.6
0.8
1.0


Joe par=2
z
x
y
0.0
0.2
0.4
0.6
0.8
1.0



2.4 Measures of Dependence

Measures of dependence are common instruments to summarize a complicated dependence structure in a single number (in bivariate sase). In this section we discuss three important measures of dependence for random vector $(X, Y)^T$. For this, we mainly follow Paul Embrechts, Filip Lindskog and Alexander McNeil (2001). Proofs an further details can be found therein.

Definition 2.12. Linear correlation is defined as

$$\rho(X,Y) \doteq \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{16}$$

where $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Unfortunately, correlation is only a suitable measure in a special class of distributions, i.e. elliptical distributions. This class includes the normal distribution and mixtures of normal distributions. It is well known, that outside this class the linear correlation coefficient is inappropriate and often misleading. The two other dependence measures to be considered are rank correlation and the coefficients of tail dependence. Both measures are general enough to give sensible measures for any dependence structure. Also they perhaps provide the best alternatives to the linear correlation coeffcient as a measure of dependence for nonelliptical distributions.

Definition 2.13. Kendall's tau is defined as

$$\tau(X,Y) = \mathbb{P}\left\{(X-\widetilde{X})(Y-\widetilde{Y}) \geq 0\right\} - \mathbb{P}\left\{(X-\widetilde{X})(Y-\widetilde{Y}) < 0\right\} \tag{17}$$

where $(\widetilde{X},\widetilde{Y})^T$ is an independent copy of $(X,Y)^T$.

It can be seem that Kendall's tau for $(X,Y)^T$ is the probability of concordance minus the probability of discordance.

Theorem 2.14. *Let Q denote the difference between the probability of concordance and discordance of $(X,Y)^T$ and $(\tilde{X},\tilde{Y})^T$, i.e*

$$Q = \mathbb{P}\left\{(X-\tilde{X})(Y-\tilde{Y}) > 0\right\} - \mathbb{P}\left\{(X-\tilde{X})(Y-\tilde{Y}) < 0\right\} \tag{18}$$

Let $(X,Y)^T$ and $(\tilde{X},\tilde{Y})^T$ be independent vectors of continuous random variables with joint distribution functions H and $\tilde{H}$, respectively, with common margins F (of X and $\tilde{X}$) and G (of Y and $\tilde{Y}$). C and $\tilde{C}$ are the copulas of $(X,Y)^T$ and $(\tilde{X},\tilde{Y})$ respectively, so that $H(x,y) = C(F(x),G(y))$ and $\tilde{H}(x,y) = \tilde{C}(F(x),F(y))$. Then

$$Q = Q(C,\tilde{C}) = 4\int\int_{[0,1]^2} \tilde{C}(u,v)dC(u,v) - 1 \tag{19}$$

See Paul(2001) for proof.

The following theorem indicates the relation between Kendall's tau for $(X,Y)^T$ and its copula C.

Theorem 2.15. *Let $(X,Y)^T$ be a vector of continuous random variables with copula C. Then Kendall's tau for $(X,Y)^T$ is given by*

$$\tau(X,Y) = Q(C,C) = 4\int\int_{[0,1]^2} C(u,v)dC(u,v) - 1 = 4\mathbb{E}(C(U,V)) - 1 \tag{20}$$

Definition 2.16. Spearman's rho is defined as

$$\rho_s(X,Y) = 3\left(\mathbb{P}\left\{(X-\tilde{X})(Y-Y') > 0\right\} - \mathbb{P}\left\{(X-\tilde{X})(Y-Y') < 0\right\}\right) \tag{21}$$

where $(X,Y)^T, (\tilde{X},\tilde{Y})^T$ and $(X',Y')^T$ are independent copies.

Theorem 2.17. *Let $(X,Y)^T$ be a vector of continuous random variables with copula C. The Spearman's rho for $(X,Y)^T$ is given by*

$$\rho_s(X,Y) = 3Q(C,\Pi) = 12\int\int_{[0,1]^2} uvdC(u,v) - 3 = 12\int\int_{[0,1]^2} C(u,v)dudv - 3 \tag{22}$$

Eq 20 & 22 will be used to get moment estimator for parametric copulas.

There is a concept that is relevant to the study of dependence between extreme values named tail dependence. We are not going deep into this topic but state some basic definitions and results.

Definition 2.18. Let $(X,Y)^T$ be a vector of continuous random variables with marginal distribution functions F and G. The coefficient of upper tail dependence of $(X,Y)^T$ is

$$\lambda_U \doteq \lim_{u\uparrow 1}\mathbb{P}\left\{Y > G^{-1}(u) \mid X > F^{-1}(u)\right\} \tag{23}$$

provided that the limit $\lambda_U \in [0,1]$ exists. If $\lambda_U \in (0,1]$, X and Y are said to be asymptotically dependent in the upper tail; if $\lambda_U = 0$, X and Y are said to be asymptotically independent in the upper tail.

It is not straightforward to see from Eq 23 that it is a concept of copula. An alternative and equivalent definition (for continuous random variables) which can be found in Joe (1997), p. 33. is the following.

Definition 2.19. If a bivariate copula C is such that

$$\lambda_U \doteq \lim_{u\uparrow 1} \frac{1-2u+C(u,u)}{1-u} \tag{24}$$

exists, them C has upper tail dependence if $\lambda_U \in (0,1]$, and upper tail independence if $\lambda_U = 0$.

Similarly the lower tail dependence can be defined.

Definition 2.20. If the limit $\lambda_L \doteq \lim_{u\downarrow 0} \frac{C(u,u)}{u}$ exiasts, then C has lower tail dependence if $\lambda_L \in (0,1]$ and lower tail independence if $\lambda_L = 0$

More discussions about tail dependence can be found in Paul Embrechts, Filip Lindskog and Alexander McNeil,(2001).

3 Copula Estimation

3.1 Parametric Estimation of Copula

Suppose that observation $\mathbf{X_1} = (X_{11},...,X_{1p})^T$,...,$\mathbf{X_n} = (X_{n1},...,X_{np})^T \overset{iid}{\sim} F_{(\theta,\eta)}$ with copula C_θ and margins $F_{1\eta},...,F_{p\eta}$ where θ is a parameter for the copula and η is a parametric for margins. Then,

$$F_{(\theta,\eta)}(x_1,..,x_p) = C_\theta(F_{1\eta}(x_1),...,F_{p\eta}(x_p). \tag{25}$$

Assume C_θ and $F_{(\theta,\eta)}$ are absolutely continuous with their densities C_θ and $f_{(\theta,\eta)}$ and marginal densities $f_{1\eta},...,f_{p\eta}$. Then,

$$f_{(\theta,\eta)}(x_1,..,x_p) = \frac{\partial^\eta C_\theta(F_{1\eta}(x_1),...,F_{p\eta}(x_p))}{\partial F_{1\eta}(x_1)...\partial F_{p\eta}(x_p)} \times f_{1\eta}(x_1) \times \cdots \times f_{p\eta}(x_p). \tag{26}$$

The likelihood of (θ,η) is

$$L(\theta,\eta) = \prod_{j=1}^{n} f_{(\theta,\eta)}(x_{j1},..,x_{jp}) \tag{27}$$

The log-likelihood is

$$ln(\theta,\eta)=\sum_{j=1}^{n}f_{(\theta,\eta)}(x_{j1},..,x_{jp}) \tag{28}$$

We can get maximum likelihood estimator $(\hat{\theta},\hat{\eta})$.

3.2 Semiparametric Estimation of Copula

- Pseudo-Likelihood

Still assume a parametric copula C_θ but the margins nonparametric since the interest here is on the dependence structure, i.e. θ. Let $F_{jn}(x)=\frac{1}{n}\sum_{l=1}^{n}I\left(X_{lj}\leq x\right)$ be the empirical distribution estimator of F_j. A pseudo (partial)-likelihood for θ is (Genest, Choahi, Riverst, 1995, Biometrika)

$$\widetilde{ln}(\theta)=\sum_{l=1}^{n}logC_\theta\left(F_{1n}(x_{l1}),...,F_{pn}(x_{lp})\right) \tag{29}$$

and $\hat{\theta}_{ps}=\underset{\theta}{argsup}\left\{\widetilde{ln}(\theta)\right\}$.

- Method of Moment

It is mostly used in the bivariate one-parameter case (see e.g. Oakes, 1982; Genest, 1987; Genest and Rivest, 1993, and the references therein). It can also sometimes be employed in the multivariate one-parameter and in the multivariate multiparameter cases. Method-of-moment approaches are based on the inversion of a consistent estimator of a moment of the copula C_θ. The two best-known moments, Spearman's rho and Kendall's tau, are respectively given by

$$\rho(\theta)=12\int_{[0,1]^2}uvdC_\theta(u,v)-3=12\int_{[0,1]^2}C_\theta(u,v)dudv-3 \tag{30}$$

$$\tau(\theta)=4\int_{[0,1]^2}C_\theta(u,v)dC_\theta(u,v)-1 \tag{31}$$

Let $\{\mathbf{R_l}\}_{l=1}^{n}$ be the vector of ranks associated with $\{\mathbf{X_l}\}_{\mathbf{l=1}}^{\mathbf{n}}$. If it is bivariate one-parameter case, consistent estimators of these two moments can be expressed as:

$$\rho_n = \frac{12}{n(n+1)(n-1)} \sum_{i=1}^{n} R_{i1} R_{i2} - 3\frac{n+1}{n-1} \tag{32}$$

$$\tau_n = \frac{4}{n(n-1)} \sum_{i \neq j} I\left[X_{i1} \leq X_{j1}\right] I\left[X_{i2} \leq X_{j2}\right] - 1 \tag{33}$$

When the functions $\rho(\theta)$ and $\tau(\theta)$ are one-to-one, consistent estimators of θ are given by $\theta_{n\rho} = \rho^{-1}(\rho_n)$ and $\theta_{n\tau} = \tau^{-1}(\tau_n)$. See Kojadinovic, I. and Yan, J. (2010) for further details about asymptotic representation of those estimators.

4 Piesewise linear approximation

In this section, we illustrated a method to study bivariate distribution using piesewise normal linear approximation.

4.1 Notations

$$\mathbf{x_i} = (x_1^i, x_2^i, ..., x_n^i) \quad i = 1, 2$$

$$\mathbf{z_i} = (z_1^i, z_2^i, ..., z_n^i) \quad i = 1, 2$$

Get $\mathbf{z_i}$ from $\mathbf{x_i}$ by: $\Phi^{-1}(\hat{F}_{\mathbf{X_i}}(x_k^i)) = z_k^i = g_i^{-1}(x_k^i)$

The empirical cdf is

$$\hat{F}_{\mathbf{X_i}}(x_k^i) = \frac{r(k) - 0.326}{n + 0.348}$$

where $r(k)$ is the rank of x_k^i and $i = 1, 2 \qquad k = 1, ..., n$

$\psi(\cdot)$ transformation from ρ_Z to ρ_X (i.e. $\psi(\rho_Z) = \rho_X$)

$\phi(\cdot)$ & $\Phi(\cdot)$ are pdf and cdf for univariate standard normal.

4.2 Truncated normal moments

According to Johnson1970, the first and second order moments of the doubly truncated variable $Z \in (a_1, a_2)$ are:

$$\mu_1(a_1, a_2) = E(Z) = \frac{\phi(a_1) - \phi(a_2)}{\Phi(a_2) - \Phi(a_1)} \tag{34}$$

$$\mu_2(a_1, a_2) = E(Z^2) = 1 + \frac{a_1\phi(a_1) - a_2\phi(a_2)}{\Phi(a_2) - \Phi(a_1)} \tag{35}$$

The probability density function (pdf) of standard binormal distribution with correlation coefficient ρ is

$$\phi(z_1, z_{2;}\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\{-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)\} \tag{36}$$

For double truncation of binormal variable (Z_1, Z_2) on a two dimentional region $D = [a_1, a_2] \times [b_1, b_2]$, the truncated pdf is $\phi(z_1, z_2; \rho)/P$, where P is the probability of D:

$$P = P(a_1, a_2, b_1, b_2; \rho) = \int_{a_1}^{a_2}\int_{b_1}^{b_2} \phi(z_1, z_2; \rho) dz_1 dz_2$$

We define $P(a, b; \rho)$ by letting a_2 and b_2 go to infinity:

$$P(a, b, \rho) = \int_{a}^{\infty}\int_{b}^{\infty} \phi(z_1, z_2; \rho) dz_1 dz_2$$

Regier and Hamdan (1971) derived the first and second order moments using Hermite polynomials, assuming single truncation on each variable Z_1 and Z_2. Dimitris and Efthymia further derived the marginal and joint first moments of $(Z_1, Z_2) \in D$ as

$$\mu_{1,0} = E(Z_1) = \frac{1}{P}\sum_{i,j=1}^{2} (-1)^{i+j}(\phi(a_i)Q(a_i, b_j; \rho) + \rho\phi(b_j)Q(b_j, a_i; \rho)) \tag{37}$$

Similarly, we can get $\mu_{0,1} = E(Z_2)$ by swapping the respective truncation points.

$$\mu_{1,1} = E(Z_1 Z_2) = \frac{1}{P} \sum_{i,j=1}^{2} (-1)^{i+j} (\rho P(a_i, b_j; \rho) + (1-\rho^2)\phi(a_i, b_j; \rho) + \rho a_i \phi(a_i) Q(a_i, b_j; \rho) + \rho b_j \phi(b_i) Q(b_i, a_j; \rho)) \quad (38)$$

In accordance with Regier and Hamdan (1971),

$$Q(a, b, \rho) = \int_{\frac{b-\rho a}{\sqrt{1-\rho^2}}}^{\infty} \phi(u) du$$

4.3 Piece-wise Linear Approximation

We consider a piece-wise linear function $g_1$with m segments decided by $m+1$ points $a_i, \quad i = 0, ..., m$. More specifically,

$$X_1 = \begin{cases} c_{10} + c_{11} Z_1 & if\ a_0 < Z_1 \leq a_1 \\ c_{20} + c_{21} Z_1 & if\ a_1 < Z_1 \leq a_2 \\ ..... & ..... \\ c_{m0} + c_{m1} Z_1 & if\ a_{m-1} < Z_1 < a_m \end{cases}$$

where $\mathbf{c_0} = [c_{10}, c_{20}, ..., c_{m0}]^T$ and $\mathbf{c_1} = [c_{11}, c_{21}, ..., c_{m1}]^T$ are the parameter vectors of constant term and slope, respectively, of the linear interpolation at the breakpoints. Note that g_1 is continuous so the solution for the coefficients are $c_{i1} = \frac{g_1(a_i) - g_1(a_{i-1})}{a_i - a_{i-1}}$ and $c_{i0} = a_i - c_{i1} g(a_i)$. Similarly, g_2 has parameter vectors $\mathbf{d_0} = [d_{10}, d_{20}, ..., d_{p0}]^T$ and $\mathbf{d_1} = [d_{11}, d_{21}, ..., d_{p1}]^T$. We define a partition $\mathcal{A}$ of the domain of X_1 by

$$\mathcal{A} = \{A_i = g_1(z_1)\ with\ z_1 \in (a_{i-1}, a_i] \mid i = 1, ..., m\}$$

The single moments of $X_1$with order k then is :

$$E(X_1^k) = \sum_{i=1}^{m} E(X_1^k \mid X_1 \in A_i) Pr(X_1 \in A_i)$$

$$E(X_1) = \sum_{i=1}^{m}(c_{i0} + c_{i1}\mu_{1;a_i})P_{a_i}$$

$$E(X_1^2) = \sum_{i=1}^{m}(c_{i0}^2 + 2c_{i0}c_{i1}\mu_{1;a_i} + c_{i1}^2\mu_{2;a_i})P_{a_i}$$

where $P_{a_i} = Pr(a_{i-1} \leq Z_1 < a_i) = \Phi(a_i) - \Phi(a_{i-1})$, $\mu_{1;a_i} = \mu_1(a_{i-1}, a_i)$ and $\mu_{2;a_i} = \mu_2(a_{i-1}, a_i)$

For the bivariate case, we denote the partition

$$\mathcal{D} = \{D_{ij} = \{a_{i-1} \leq z_1 \leq a_i, b_{j-1} \leq z_2 \leq b_j\} \mid i = 1, ..., m \;\; j = 1, ..., p\}$$

Applying the total probability law in Eq. 8 the first order joint moment of (X_1, X_2) is

$$E(X_1, X_2) = \sum_{i=1}^{m}\sum_{j=1}^{p} E(X_1X_2 \mid X_1 \in A_i \;\wedge\; X \in B_j)Pr(X_1 \in A_i \;\wedge\; X \in B_j)$$

Notice here

$$X_1 \mid A_i = c_{i0} + c_{i1}Z_1 \qquad a_{i-1} < Z_1 \leq a_i$$

$$X_2 \mid B_j = d_{j0} + d_{j1}Z_2 \qquad b_{j-1} < Z_2 \leq b_j$$

So we can get

$$E(X_1X_2) = \sum_{i=1}^{m}\sum_{j=1}^{p}(c_{i0}d_{j0} + c_{i1}d_{j0}\mu_{1,0} + c_{i0}d_{j1}\mu_{0,1} + c_{i1}d_{j1}\mu_{1,1})P$$

Based on all the reaulte above we can calculate ρ_X based on piese-wise linear transformation by the following:

$$\rho_X = \psi(\rho_Z) = \frac{cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}} = \frac{E(X_1X_2) - E(X_1)E(X_2)}{\sqrt{EX_1^2 - (EX_1)^2}\sqrt{EX_2^2 - (EX_2)^2}}$$

4.3.1 Estimation of the normal correlation coefficient

Now we seek the story in another direction that is to solution for ρ_Z from two non-normal variables X_1 and X_2. Suppose $(Z_1, Z_2) \sim (0, 0, 1, 1, \rho_Z)$ and (X_1, X_2) with marginal cdf $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ and with correlation

coefficient ρ_X. The cdf and be approximated by emperical cdf based on observations $\mathbf{x_1}$ and $\mathbf{x_2}$. The transformation g_1 and g_2 are as defined before. It was shown in Cario and Nelson (1996) that the correlation transform $\rho_X = \psi(\rho_Z)$ is also monotonic and therefore a solution $\rho_Z = \psi^{-1}(\rho_X)$ can be found. Based on the form of ψ, an interative scheme is provided by D. Kugiumtzis and E. Bora-Senta (2010).

1. Set the breakpoints and compute the coefficients $\mathbf{c_0}, \mathbf{c_1}, \mathbf{d_0}, \mathbf{d_1}$ of the piece-wise linear approximation of g_1 and g_2 as defined in Section 3 which determine the form of $\psi(\rho_Z)$.

2. Begin with a starting value $\rho_{Z,0} \in (-1, 1)$.

3. At each iteration i, compute $\rho_{X,i} = \psi(\rho_{Z,i})$ and the difference $\delta_{\rho_{X,i}} = \rho_X - \rho_{X,i}$.

4. If $\delta_{\rho_{X,i}} < \epsilon$ the solution is found as $\rho_Z = \rho_{Z,i}$ where ϵ is an arbitary tolerance. Otherwise update the approximation of ρ_Z as

$$\rho_{Z,i+1} = \begin{cases} \rho_{Z,i} + \delta_{\rho_{X,i}} & if \ \mid \rho_{Z,i} + \delta_{\rho_{X,i}} \mid < 1 \\ sgn(\rho_{Z,i}) min(\frac{|\rho_{Z,i}|+1}{2}, 2 - \mid \rho_{Z,i} + \delta_{\rho_{X,i}} \mid) & if \ \mid \rho_{Z,i} + \delta_{\rho_{X,i}} \mid \geq 1 \end{cases} \tag{39}$$

   where $sgn(x)$ is the sign of x . Then go to step 3.

4.3.2 Generation of linearly correlated bivariate samples

Algorithm from D. Kugiumtzis and E. Bora-Senta (2010).

1. Cumpute the sample correlation coefficient r_X and marginal cdf $\hat{F}_{X_1}(x_1)$ and $\hat{F}_{X_2}(x_2)$ of (X_1, X_2) from the sample $(\mathbf{x_1}, \ \mathbf{x_2})$

2. Compute the normal correlation coefficient r_Z of the corresponding bivariate standard normal (Z_1, Z_2) by D. Kugiumtzis and E. Bora-Senta's algorithm.

3. Generate a sample $(\mathbf{z_1}, \ \mathbf{z_2})$ of (Z_1, Z_2), i.e. a sample of size n from a bivariate standard normal distribution with correlation coefficient ρ_Z.

4. Transform $(\mathbf{z_1}, \mathbf{z_2})$ to $(\mathbf{x_1^*},\ \mathbf{x_2^*})$ as

$$\mathbf{x_i^*} = \hat{F}_{X_1}^{-1}(\Phi(\mathbf{z_i}))$$

5 Apply to nutrition measurement data

5.1 Data description

In this section, we explore the relationship bwtween two dietary intake variables using methods talked before. The two variables are energy intake and vitamin C intake. There are 571 observations for each.

We choose to be $\rho_{Z0} = cor(x1, x2)$.

5.2 Dependence structure exploration

We estimated the copula with both parametric and non-parametric method. Two parametric copulas, t copula and normal copuls are applied. Other than that, the following two non-parametric methods are also used.

- Empirical estimator proposed by Deheuvels (1979):

$$\widetilde{C}(u, v) = \frac{1}{n} \sum_{i=1}^{n} I\left(\hat{U}_i \leq u, \hat{V}_i \leq v\right) \tag{40}$$

where $\hat{U}_i = \frac{1}{n} \sum_{j=1}^{n} I\left(X_{j1} \leq X_{i1}\right)$ and $\hat{V}_i = \frac{1}{n} \sum_{j=1}^{n} I\left(X_{j2} \leq X_{i2}\right)$.

- A two-stage kernel estimator (Chen and Huang 2007)

As copulas are not directly observable, a nonparametric copula estimator has to be formed in two stages: estimate the two marginals $(F_1(X_1), F(X_2))$ first and then estimate the copula based on the estimated marginals. Let K be a symmetric probability density supported on $[-1, 1]$ and $G(x) = \int_{-\infty}^{x} K(t)dt$ be the distribution of K. In the first stage the marginal distribution F_l is estimated by

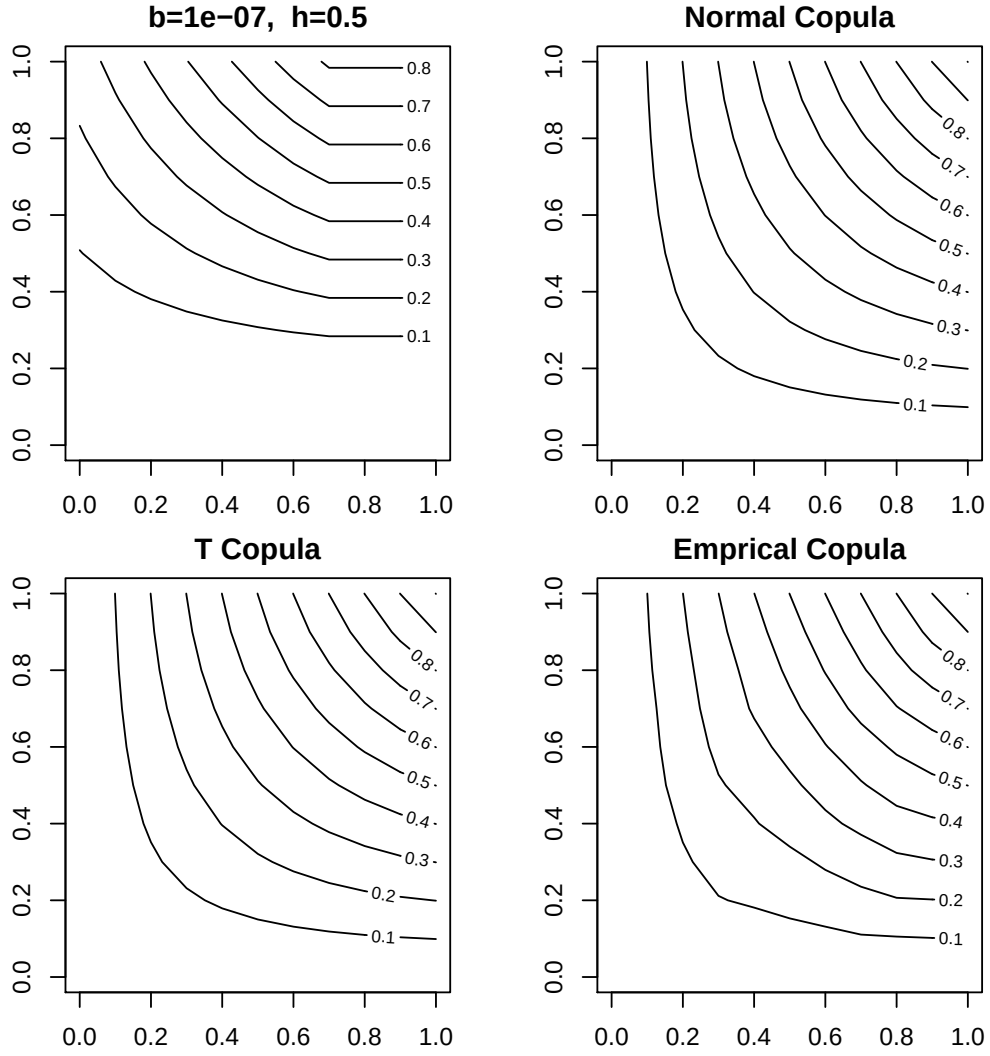
$$\hat{F}_l(x) = \frac{1}{n} \sum_{i=1}^{n} G\left\{\frac{x - X_{il}}{b_l}\right\} \tag{41}$$

with a bandwidth b_l for $l = 0, 1, 2$; see Bowman, Hall & Prvan (1998) for details on this kernel distribution estimator. A estimator of $C(u, v)$ is

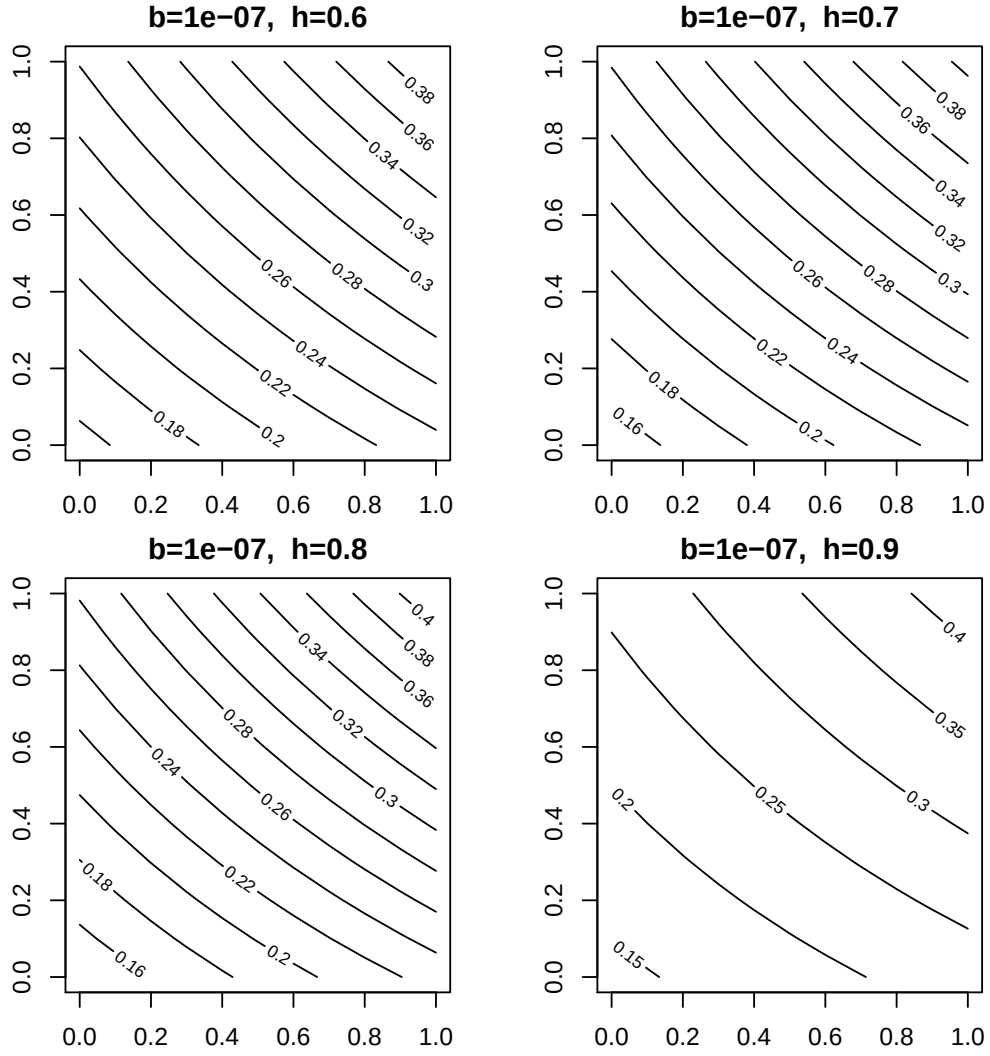
$$\frac{1}{n}\sum_{i=1}^{n} G\left(\frac{u - \hat{F}_1(X_{i1})}{h}\right) G\left(\frac{v - \hat{F}_2(X_{i2})}{h}\right). \tag{42}$$

The estimator considered by Chen (2007) is of above form based on the local linear kernels $K_{u,h}$ which can be used to prevent the boundary bias.

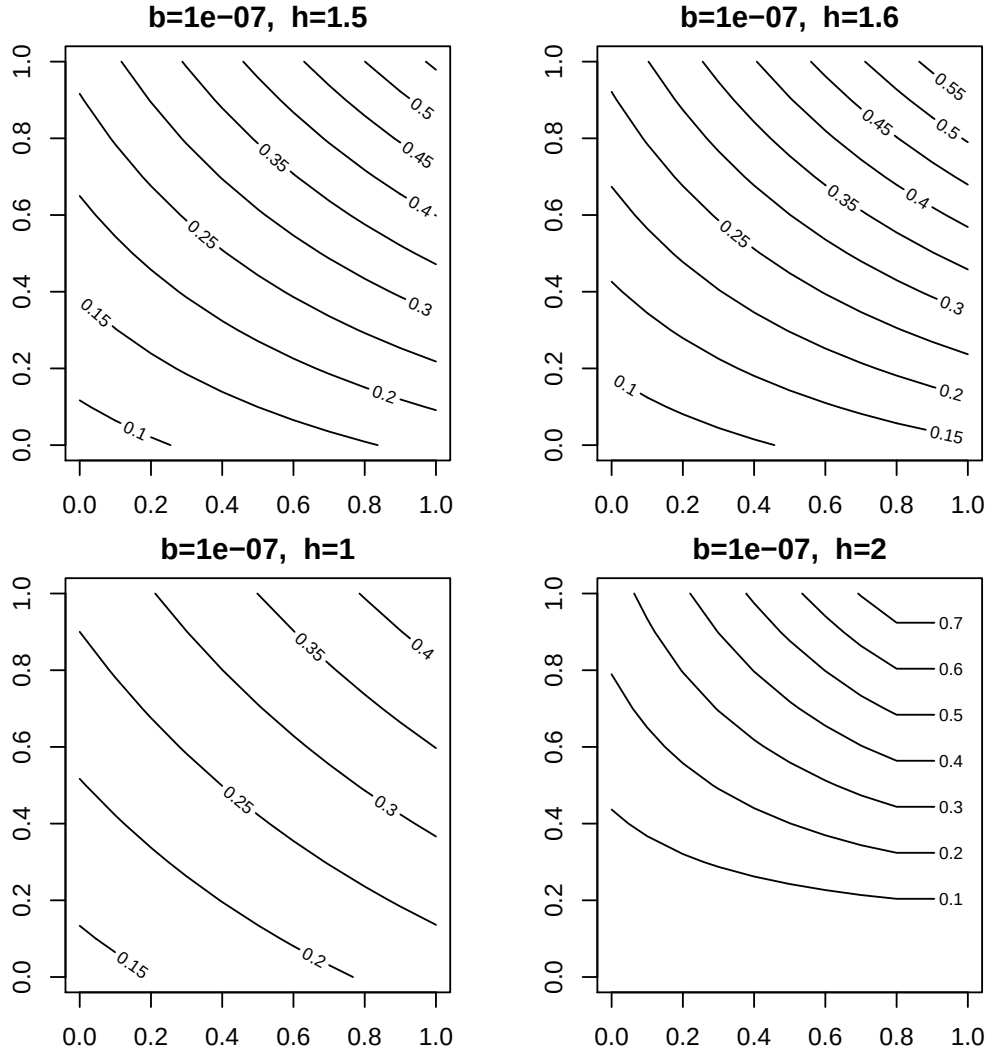
```
> par(mfrow=c(2,2),mar=c(1,2,1,1)+1)
> contour(u,v,z0.5,main="b=1e-07,  h=0.5")
> contour(u,v,zga,main="Normal Copula")
> contour(u,v,zt,main="T Copula")
> contour(u,v,empest,main="Emprical Copula")
```




```
> par(mfrow=c(2,2),mar=c(1,2,1,1)+1)
> contour(u,v,z0.6,main="b=1e-07,  h=0.6")
> contour(u,v,z0.7,main="b=1e-07,  h=0.7")
> contour(u,v,z0.8,main="b=1e-07,  h=0.8" )
> contour(u,v,z0.9,main="b=1e-07,  h=0.9")
```




```
> par(mfrow=c(2,2),mar=c(1,2,1,1)+1)
> contour(u,v,z1.5,main="b=1e-07,  h=1.5")
> contour(u,v,z1.6,main="b=1e-07,  h=1.6")
> contour(u,v,z1.0,main="b=1e-07,  h=1")
> contour(u,v,z2.0,main="b=1e-07,  h=2")
```



5.3 Model fitting

We fit copula and piecewise linear approximation models separately with the nutrition measurement data. The estimated results are for copula are:

```
[1] "Pseudo-Likelihood estimator for t copula"

       rho         nu
 0.2753902 18.5563999
```

```
[1] "inversion of Kendall's tau"

       Estimate Std. Error  z value     Pr(>|z|)
rho.1 0.2759493 0.04059246 6.798044 1.060494e-11
```

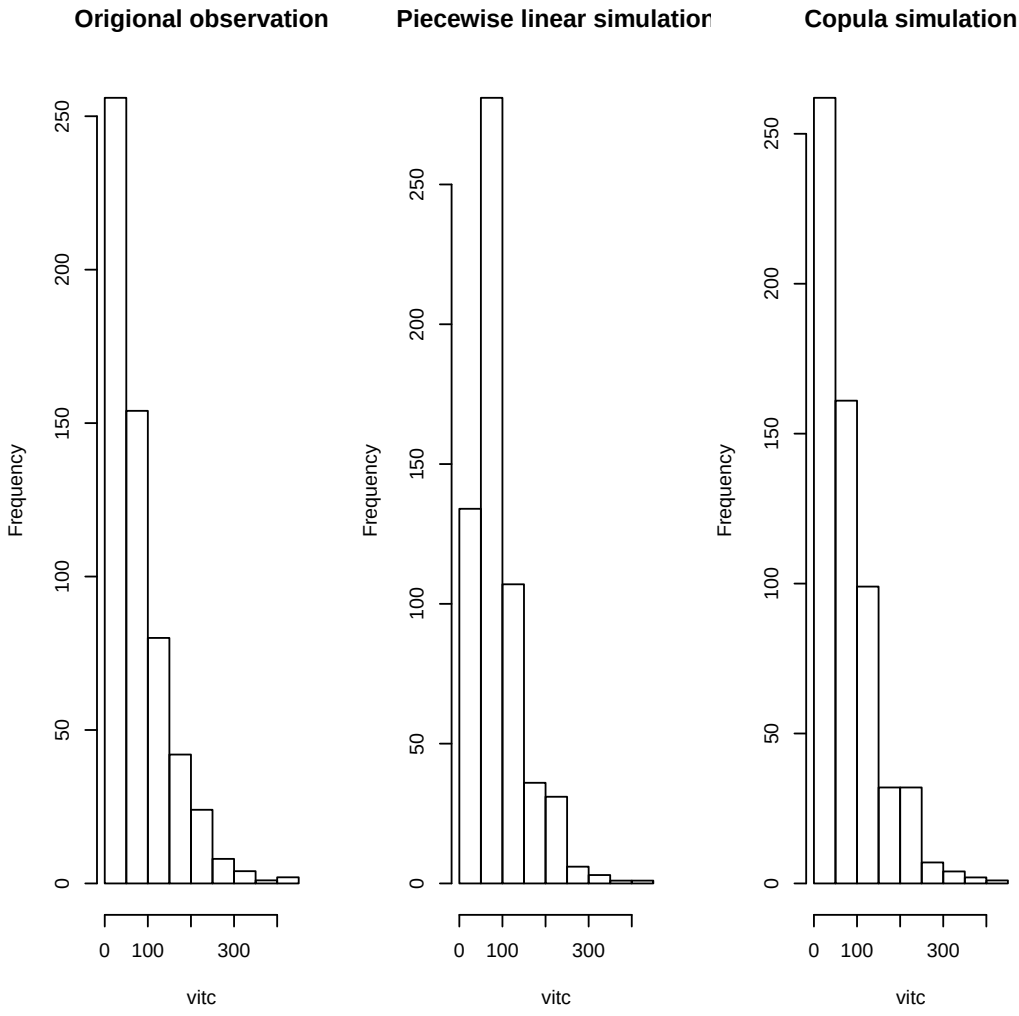
```
> rbind(c0,c1)

       [,1]     [,2]      [,3]        [,4]     [,5]      [,6]
c0 68.26921 52.97509  25.52315  -0.5740349 -71.5966 -837.2283
c1 22.22071 91.37781 118.44066 134.4870654 167.2871  461.9189
```

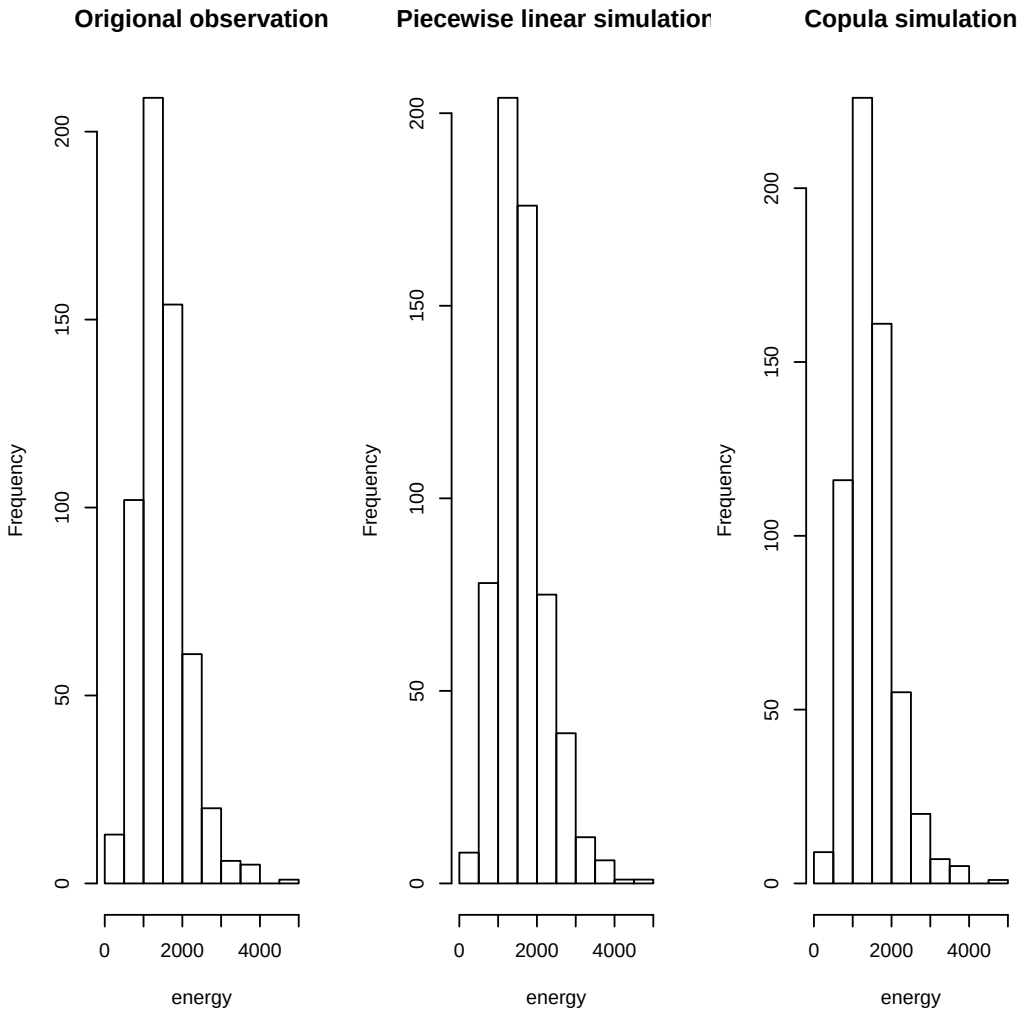
```
> rbind(d0,d1)

        [,1]      [,2]      [,3]
d0 1463.8712 1345.5853 -139.2652
d1  443.9166  834.4287 1547.6951
```

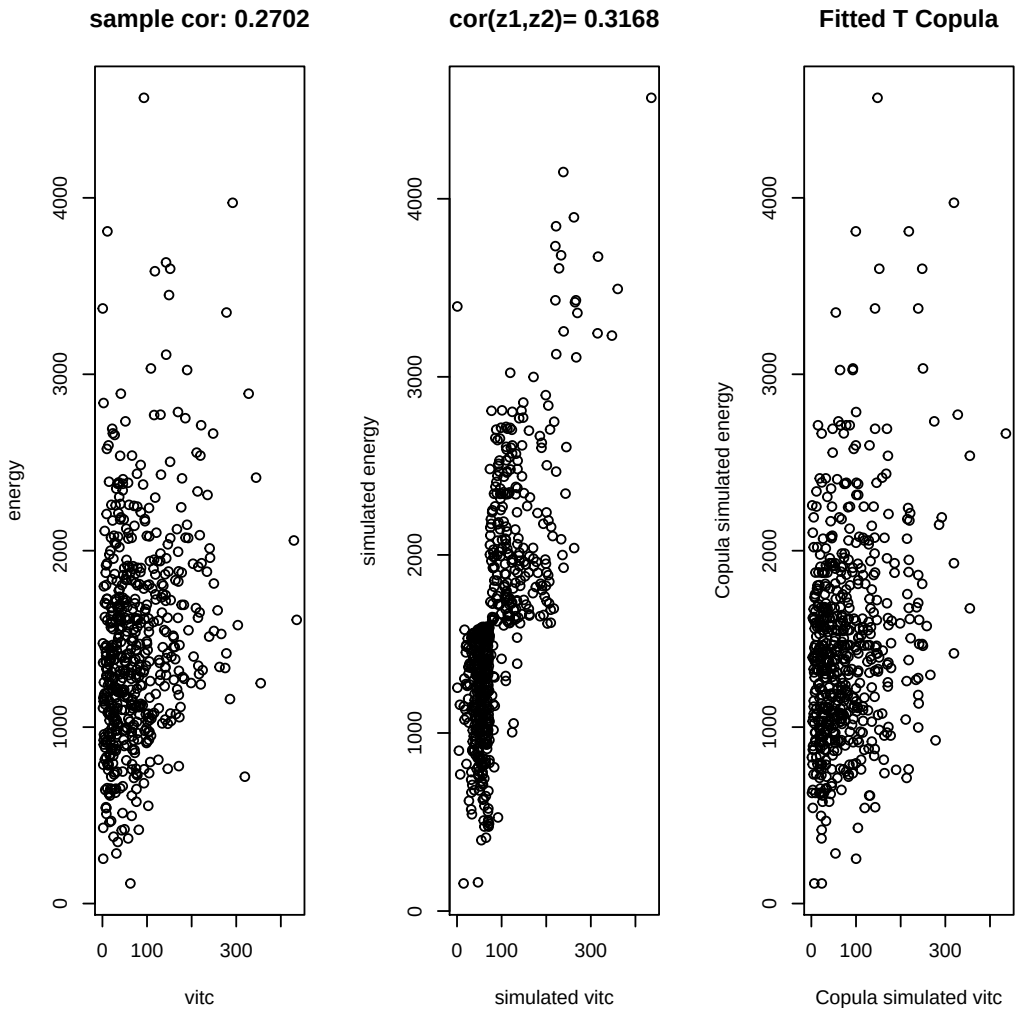
5.4 Generation of bivariate random samples

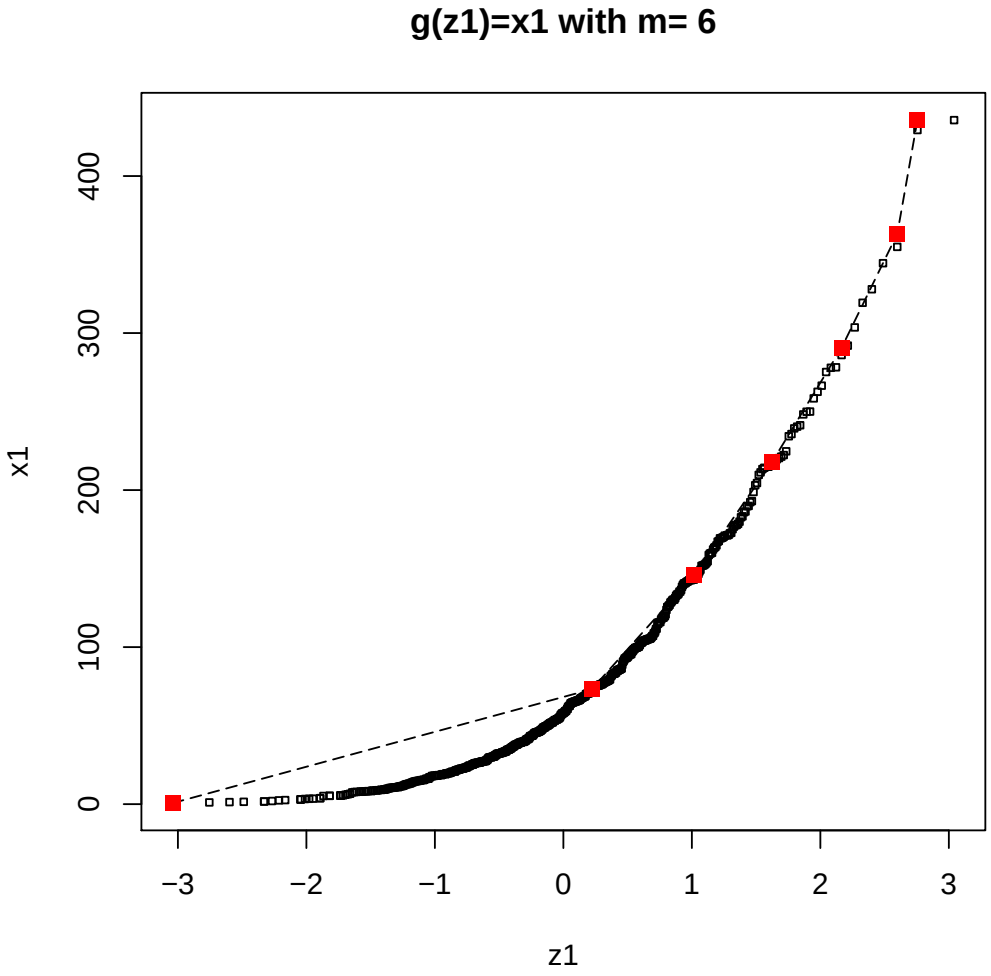
The summary statistics based on simulated vitc and energy data are shown below.

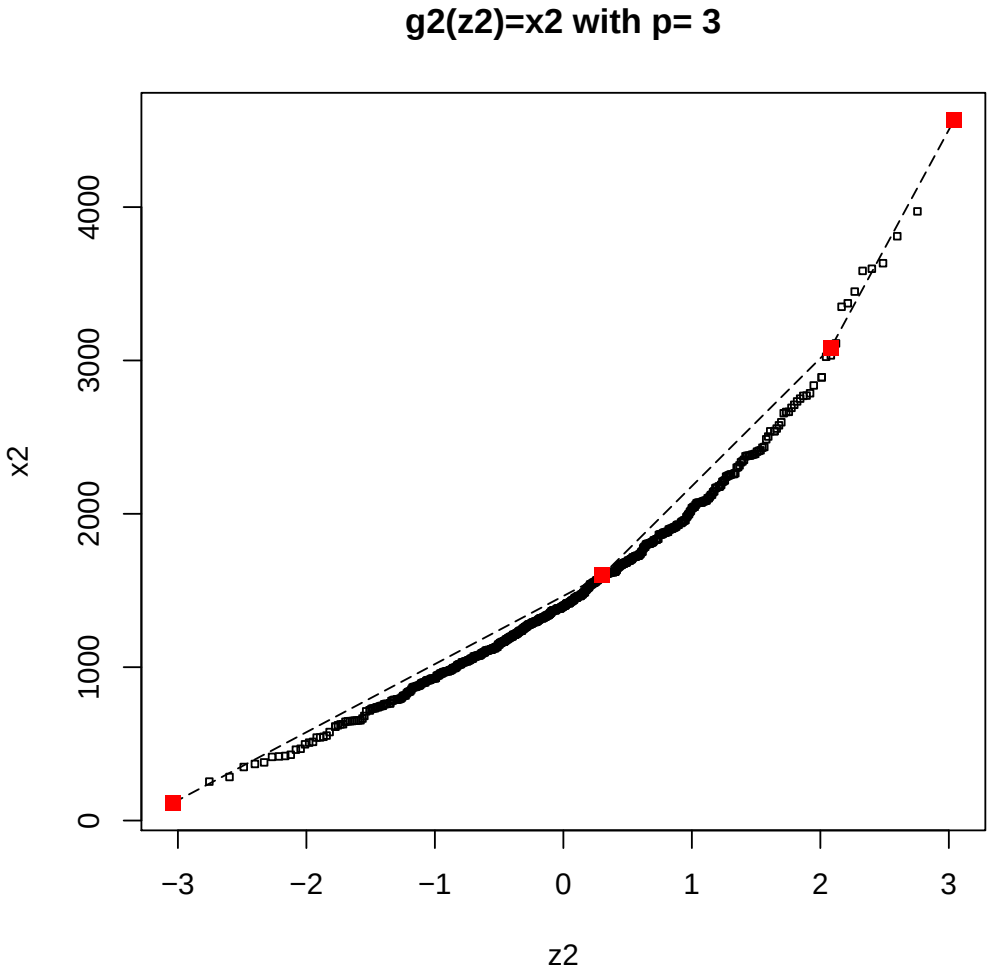
```
                   Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
origional_vic      0.70   26.30   58.10   77.92   105.4  435.6
piecewise_vic      0.70   51.60   69.55   89.64   112.8  435.6
copula_vic         1.31   27.45   59.45   78.91   107.5  435.6
origional_energy 114.00 1074.00 1399.00 1482.00  1806.0 4567.0
piecewise_energy 155.10 1161.00 1515.00 1607.00  1952.0 4567.0
copula_energy    114.00 1054.00 1402.00 1465.00  1743.0 4567.0
```


Origional observation
Piecewise linear simulation
Copula simulation
Frequency
0
50
100
150
200
250
0
100
300
vitc


Origional observation
Piecewise linear simulation
Copula simulation
Frequency
0
50
100
150
200
0
2000
4000
energy


sample cor: 0.2702
cor(z1,z2)= 0.3168
Fitted T Copula
energy
simulated energy
Copula simulated energy
vitc
simulated vitc
Copula simulated vitc
0
1000
2000
3000
4000
0
100
300


g(z1)=x1 with m= 6
x1
z1
0
100
200
300
400
−3
−2
−1
0
1
2
3


g2(z2)=x2 with p= 3
x2
0
1000
2000
3000
4000
−3
−2
−1
0
1
2
3
z2

References

[1] M. Frechet. (1951) Sur les tableaux de correlation dont les marges sont donnes, Ann. Univ. Lyon, Science, 4, 13-84

[2] Frechet, M. R. (1958). Remarques au sujet de la note precedente. C. R. Acad. Sci. Paris Ser. I Math. 246, 2719-2720.

[3] G. Dall, Aglio. (1972) Frechet classes and compatibility of distribution functions, Symposia Math., 9, 131-150

[4] Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In Dempster, M., editor, Risk Management: Value at Risk and Beyond., pages 176-223. Cambridge Univ. Press, Cambridge.

[5] Li, D. (2001): On default correlation: a Copula function approach, Journal of Fixed Income,9, 43-54.

[6] Yu GH, Huang CC (2001) A distribution free plotting position. Stoch Environ Res Risk Assess 15:462-476

[7] Johnson N, Kotz S (1970) Distributions in statistics, continuous univariate distributions. Houghton Mifflin Company, Boston

[8] Regier MH, Hamdan MA (1971) Correlation in a bivariate normal distribution with truncation in both variables. Aust J Stat 13(2):77-82

[9] Cario MC, Nelson BL (1996) Autoregressive to anything:time-series inpute processes for simulation. Oper Res Lett 19:51-58

[10] Dimitris Kugiumtzis, Efthymia Bora-Senta (2010) Normal correlation coefficient of non-normal variables using piece-wise linear approximation, Comput Stat 25:645-662

[11] La fonction de dependance empirique et ses proprietes. Un test non paramerique d'independance. Acad. Roy. Belg. Bull. Cl. Sci., 65, 274-292.

[12] S.X.Chen & T.M.Huang. (2007). "Nonparametric estimation of copula functions for dependence modeling", The Canadian Journal of Statistics, Vol 35, No. 2, 265-282.

[13] A. Bowman, P. Hall & T. Prvan (1998). Bandwidth selection for the smoothing of distribution functions. Biometrika, 85, 799-808

[14] Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82, 543-552

[15] Oakes,D., (1982). A model for association in bivariate survival data, Journal of the Royal Statistical Society Series B 44, 414-422.

[16] Genest, C.,(1987). Frank's family of bivariate distributions, Biometrika 74 (3), 549-555.

[17] Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association, 88 (423), 1034-1043.

[18] Kojadinovic, I. and Yan,J. (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. Insurance:Mathematics and Economics 47, 52-63

[19] Paul Embrechts, Filip Lindskog and Alexander McNeil,(2001), Modelling Dependence with Copulas and Applications to Risk Management, Working paper, ETH, Zurich, http://www.math.ethz.ch/Finance

[20] T. Schmidt, (2007), Coping with copulas, J. Rank (Ed.), Copulas-From Theory to Application in Finance, Risk Books, London, pp. 3-34

[21] Sklar, A. (1959). Fonctions de repartition a n dimensions e leurs marges. Publications de l'Institut de Statistique de l'Univiversite de Paris 8, 229-231.

[22] Hoeffding, W. (1940). Scale-invariant correlation theory. In N. I. Fisher and P. K. Sen (Eds.),The Collected Works of Wassily Hoeffding, pp.57-107. New York:Springer-Verlag.

[23] Mikusinski, P., H. Sherwood, and M. Taylor (1992), The Frechet bounds revis-ited, Real Analysis Exchange, 17, 759-764.

[24] Joe, H. (1997): Multivariate Models and Dependence Concepts. Chapman & Hall, London.

[25] Nelsen, R. (1999): An Introduction to Copulas. Springer, New York.