American Journal of Mathematics and Statistics 2019, 9(4): 160-164
DOI: 10.5923/j.ajms.20190904.02

Copula Density Estimation of Iranian Household Income and Expenditure by Using Selection Method

S. Shams*, H. Rashidi, S. Rezaee

Department of Statistics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Abstract Recently by using contamination families, a new way of modeling dependence has been introduced. In this method, a sequence of parametric copulas is considered and in a few numbers of steps, accurate approximations for copula densities are obtained. By using the selection model method, the model complexity and number of model parameters are balanced. In this paper, two main variables in Iranian Household Income and Expenditure survey are considered and a copula density for those variables is estimated by using contamination family and selection model method.

Keywords Contamination family, Copula density, Fourier coefficients, Household Income and Expenditure Survey, Legendre polynomials, Selection model

1. Introduction

In multivariate studies, measures of dependence that are invariant under special transformations are too important. Also, the linear correlation has many restrictions in applications, Embrechts et al. (2003) and Mc Neil et al. (2005) considered other forms of correlations. The copula approach is a useful method for separating univariate margins and the multivariate dependence structure by using Sklar's theorem (1959, 1996). Nelsen (2006) drew attention to copula distribution function and dependence.

The problem of copula density estimation has been studied in Biau and Wegkamp (2005), and this subject has been developed by Kallenberg (2008) by using exponential families and contamination families.

Kallenberg (2009) focused on estimating the (unknown) copula density by the selection method. In this method, the modeling step consists of an intermediate approach between a parametric family and a non-parametric approach. This is done by considering a sequence of parametric copula models and starting with a given copula density or a given family of copula densities. In order to balance between the complexity of the model and the number of parameters, the model selection techniques determine which aspects are the most important ones to capture into our model.

This paper is organized as follows. Section 2 deals with some preliminaries. In section 3 the exponential families are reviewed and the decomposition of the total error into the model error and the stochastic error is explained. In Section 4 the contamination families based on Legendre polynomials are reviewed, also this section deals with the model selection problem to choose the best dimension with fast convergence to probability 1. In section 5, for two main variables, Income and Expenditure, in Iranian Household Income and Expenditure Survey, the nearest approximation of copula density using the selection method is obtained.

* Corresponding author:
s.shams@alzahra.ac.ir (S. Shams)
Published online at http://journal.sapub.org/ajms

2. Preliminaries

A 2-dimensional **copula** is a function $C : [0,1]^2 \rightarrow [0,1]$ with the following properties:

1) For every $u, v \in [0, 1]$, $C(0, v) = C(u,0) = 0$;
2) For every $u, v \in [0,1]$, $C(u. 1) = u$, $C(1,v) = v$;
3) For every (u_1, v_1), $(u_2, v_2) \in [0, 1] \times [0, 1]$ with
$u_1 \leq u_2, v_1 \leq v_2$;
$C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$

The theoretical basis of multivariate modeling by copulas is provided by a theorem due to Sklar (1959), known as **Sklar's Theorem**. Let F be a joint distribution function with margins F_1, F_2 which are respectively the cumulative distribution functions of the random variables X_1 and X_2. Then there exists a copula function C such that

$$F(x_1,x_2) = C(F_1(x_1), F_2(x_2)) \quad (1)$$

For every $x_1, x_2 \in \bar{R}$ where $\bar{R}$ represents the extended real line. Conversely, if C is a copula and F_1, F_2 are distribution functions then the function F defined a joint distribution function with margins F_1, F_2.

The parametric copula approach ensures a high level of flexibility for modeling, because the dependence structure

can be separated from the margins, through the function C with an underlying parameter θ which governs the intensity of the dependence.

In the case that the bivariate distribution has a density f, and this is available, we have

$$f(x_1,x_2) = c(F_1(x_1), F_2(x_2)).f_1(x_1).f_2(x_2) \quad (2)$$

where c is the copula density and it should be approximated in most cases.

In general, a natural and very useful way to describe a smooth function on the interval (0, 1) is to apply the orthonormal system of Legendre polynomials. This leads to a function z on (0, 1) as

$$z(u) = \sum_{r\geq 0} \gamma_r\, b_r(u) \quad (3)$$

where b_r is the r^{th} Legendre polynomial on (0, 1) and γ_r is the r^{th} Fourier coefficient, such that

$$\gamma_r \leq z, b_r \geq \int_0^1 z(u)b_r(u)du \quad (4)$$

For example, the Legendre polynomials b_0, ..., b_5 are given by

$$\begin{aligned}
b_0(u) &= 1\\
b_1(u) &= \sqrt{3}(2u-1)\\
b_{r2}(u) &= \sqrt{5}(6u^2-6u+1)\\
b_3(u) &= \sqrt{7}(20u^3-30u^2+12u-1)\\
b_4(u) &= 3(70u^4-140u^3+90u^2-20u+1)\\
b_5(u) &= \sqrt{11}(252u^5-630u^4+560u^3-210u^2+30u-1)
\end{aligned} \quad (5)$$

In the next section, it is seen that by using Legendre Polynomials a copula density function (given a known starting copula density function) is approximated.

3. Exponential Families

The exponential families are well-known families of parametric models that are used for approximating copula density function. If c_0 is the starting copula density function, the desired copula density is then approximated by

$$c_k(u, v; \theta) = c_0(u, v)\exp\{\sum_{j=1}^k \theta_j h_j(u, v) - \psi_k(\theta)\} \quad (6)$$

where $h_j(u, v) = b_{rj}\, b_{sj}$, b_{rj} and b_{sj} are Legendre polynomials, $\theta = (\theta_1, \ldots, \theta_k)$ is the vector of parameters and ψ_k is a normalizing function given by

$$\psi_k(\theta) = \log \int\int c_0(u, v)\exp\{\sum_{j=1}^k \theta_j h_j(u, v)\}dudv \quad (7)$$

Obviously, increasing the number of parameters yields to model with more complexity, so in order to balance between complexity and the number of parameters, dimension k is determined. Note that c_0 may contain unknown parameters, which should be estimated as well.

Equation (2) shows that $\log(\frac{c_k}{c_0})$ is approximated by a linear combination of the functions h_j minus a normalizing factor ψ_k (to make its integral equal to 1). Exponential families ensure automatically that we get densities such that θ belongs to the natural parameter space

$$\Theta = \{\theta; \int\int c_0(u, v)\exp\{\sum_{j=1}^k \theta_j h_j(u, v)\}dudv < \infty\} \quad (8)$$

The criteria for choosing the best approximation might be the Kullback Leibler information, $K(c,c_k(\theta))$, given by

$$\begin{aligned}
K(c,c_k(\theta)) &= E_c \log\left(\frac{c}{c_k(\theta)}\right)\\
&= E_c \log c - E_c \log(c_k(\theta))\\
&= E_c \log c/c_0 - \{\sum_{j=1}^k \theta_j E_c h_j - \psi_k(\theta)\}\\
&= K(c,c_0) - \{\sum_{j=1}^k \theta_j E_c h_j - \psi_k(\theta)\}\\
&= K(c,c_0) - K(c_k(\theta),c_0)\\
&\quad + \sum_{j=1}^k \theta_j(E_\theta h_j - E_c h_j).
\end{aligned} \quad (9)$$

It is seen that minimizing $K(c_0,c_k(\theta))$ is equivalent to maximizing $\sum_{j=1}^k \theta_j E_c h_j - \psi_k(\theta)$, which gives the asymptotic version of the maximum likelihood estimator. So, asymptotically the maximum likelihood estimator chooses that member $c_k(\theta)$ of the exponential family which is closest to the true density c in terms of Kullback Leibler information criteria.

Kallenberg (2008) showed that $c_k(\tilde{\theta})$ is the projection of c into the exponential family with base c_0, because

$$K(c,c_0) = K(c,c_k(\tilde{\theta})) + K(c_k(\tilde{\theta}),c_0) \quad (10)$$

$$K(c,c_k(\tilde{\theta})) = min\{K(c,c_k(\theta));\ \theta \in \Theta\} \quad (11)$$

Where $\tilde{\theta} \in int\Theta$ is a unique point such that

Hence $K(c,c_0)$, as the model error, is reduced to $K(c,c_k(\tilde{\theta}))$, with a reduction equal to $K(c_k(\tilde{\theta}),c_0)$. Another extra reduction from taking a higher dimension, when going from k to k+1, is occurred by an amount $K(c_{k+1}(\tilde{\theta}_{k+1}), c_0) - K(c_k(\tilde{\theta}_k, c_0)$. For the exponential family, the better fit means the smaller model error and the higher dimension or the more parameters have to be estimated. Since parameters estimation in the exponential family is difficult, the idea of contamination family is developed.

4. Contamination Families

As mentioned in Kallenberg (2009), just like the exponential family, the starting point is a copula density c_0, and $c - c_0$ is approximated by a linear combination of the functions $b_r(U)$, $b_S(V)$, hence

$$c_k(u,v) - c_0(u,v) = \sum_{j=1}^k \gamma_{r_j s_j} b_{r_j}(u) b_{s_j}(v) \quad (12)$$

where γ_{rs} are Fourier coefficients as follows

$$\begin{aligned}\gamma_{rs} &= \int\int\{c\ (u,v) - c_0\ (u,v)\}\ b_r(u)\ b_S(v)\ dudv \\ &= E_C(b_r(U)b_S(V)) - E_{c_0}(b_r(U)b_S(V)) \\ &= \rho(b_r(U),b_S(V);\ c) - \rho(b_r(U),b_S(V);\ c_0)\end{aligned} \tag{13}$$

These coefficients depend on the unknown copula density function c that if it is replaced with empirical copula mass function c_n, then γ_{rs} can be estimated as

$$\hat{\gamma}_{rs} = \frac{1}{n}\sum_{i=1}^{n} b_r(U_i)b_s(V_i) - E_{c_0}(b_r(U)b_s(V)) \tag{14}$$

Again when the starting copula density function c_0 belongs to a parametric family, its parameters should be estimated, then we have

$$\hat{c}_k\ (u,v) = c_0\ (u,v)\ + \sum_{j=1}^{k} \hat{\gamma}_{r_j s_j} b_{r_j}(u) b_{s_j}(v) \tag{15}$$

Kallenberg (2009) showed that by considering the term $\| c - \hat{c}_k\ (\theta)\|_2^2$ as the model error given by

$$\begin{aligned}\| c - \hat{c}_k \|_2^2 &= \| c - c_k \|_2^2 + \| c_k - \hat{c}_k \|_2^2 \\ &= \left(\sum_{r,s} \gamma_{rs}^2 - \sum_{j=1}^{k} \gamma_{r_j s_j}^2\right) + \sum_{j=1}^{k} (\gamma_{r_j s_j} - \hat{\gamma}_{r_j s_j})^2.\end{aligned} \tag{16}$$

where $// f \|_2^2 = \int\int (f(u,v))^2\, dudv$.

By equation (16) it can be seen that Total Error is decomposed by Model Error and Stochastic Error,

Total Error = Model Error + Stochastic Error

The model error $// c - c_k) \|_2^2$ expresses how good the contamination family approximates the true density c and the stochastic error $\| c_k - \hat{c}_k \|_2^2$ is due to estimation.

4.1. Model Selection

In order to obtain parameter estimations in contamination family, the best dimension should be chosen. Suppose m_n be the largest dimension of r and s with n observations, then we have

$$\hat{c}_{m_n}(u,v) = c_0(u,v) + \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} \hat{\gamma}_{r_j} b_r(u) b_s(v) \tag{17}$$

For the selection rule, taking all the coefficients $\hat{\gamma}_{rs}$ for $1 \le r, s \le m_n$ yields a large estimation error, so we consider only the largest Fourier coefficients and ignore the rest. Therefore, the estimator from (17), is replaced by restricting to the k_n largest among $\hat{\gamma}_{rs}$ with $1 \le r, s \le m_n$, yielding

$$\hat{c}(u,v) = c_0(u,v) + \sum_{j=1}^{k_n} \hat{\gamma}_{r_j s_j} b_{r_j}(u) b_{s_j}(v) \tag{18}$$

With

$$|\hat{\gamma}_{r_1 s_1}| \ge |\hat{\gamma}_{r_2 s_2}| \ge \cdots \ge |\hat{\gamma}_{r_{K_n} s_{K_n}}| \tag{19}$$

Random variables r_j and s_j depend on the data, and they are not chosen in advance. So how large should we take k_n? The optimal choice depends on c, but c is unknown, hence a data-driven selection of the dimension is taken.

From (9), the model error for $\hat{c}_k$ is $\sum_{r,s} \gamma_{rs}^2 - \sum_{j=1}^{k} \gamma_{r_j s_j}^2$. Hence, $\sum_{j=1}^{k} \gamma_{r_j s_j}^2$ should grow sufficiently fast in order to take a higher dimension. For that purpose a penalty is introduced, classical penalties are for example $n^{-1}\log n$ (Schwarz's rule) or $2n^{-1}$ (Akaike's criterion). It may be better to take a larger penalty, taking into account the variance of $\hat{\gamma}_{rs}{}^2$.

Kallenberg (2009), introduced a penalty as

$$\Delta_n = n^{-1}(\log n)(\log m_n) \tag{20}$$

And the selection rule as:

$$\hat{k} = \begin{cases} 0 & (\hat{\gamma}_{r_1 s_1})^2 < \Delta_n \\ \max\{1 \le k \le m_n; (\hat{\gamma}_{r_k s_k})^2 \ge \Delta_n\} & otherwise \end{cases} \tag{21}$$

The estimated copula density now becomes

$$\hat{c}(u,v) = c_0(u,v) + \sum_{j=1}^{\hat{k}} \hat{\gamma}_{r_j s_j} b_{r_j}(u) b_{s_j}(v) \tag{22}$$

5. Copula Density Estimation for Iranian Household Income and Expenditure

The 2015 *IHIE* survey was carried out by a sample of 18839 households in urban areas and 19340 households in rural areas. The survey target population includes all private and collective settled households in urban and rural areas. *A* three-stage cluster sampling method with strata is used in the survey. At the first stage, the census areas are classified and selected. At the second stage, the urban and rural blocks are selected and the selection of sample households is done at the third stage. The number of samples is optimized to estimate average annual income and expenditure of the sample household based on the aim of the survey. In this section, the model selection method is used to estimate copula density for Iranian Household Income and Expenditure (*IHIE*). Income and Expenditure descriptive statistics of urban and rural household are shown in Tables 1 and 2, respectively.

Table 1. Descriptive statistics for Income and Expenditure data of Urban household

Descriptive Statistics	Income (10^6Rials)	Expenditure (10^6Rials)
Minimum	4.914	1.400
Q_1	141.30	124.41
Median	207.43	182.18
Weighted Mean	278.78	262.87
Q_3	299.27	270.61
Maximum	5787	4424

Table 2. Descriptive statistics for Income and Expenditure data of Rural household

Descriptive Statistics	Income (10^6Rials)	Expenditure (10^6Rials)
Minimum	3.6	1.23
Q_1	84.52	79.41
Median	136.81	124.77
Weighted Mean	161.19	147.26
Q_3	205.58	185.43
Maximum	5743	3876

5.1. IHIE Copula Density Estimation with Contamination Families

By using empirical distributions as marginal distribution estimations for both variables as

$$F_n^x(x) = (n+1)^{-1}\sum_{i=1}^{n} 1(X_i \le x) \tag{23}$$

Now the problem is to estimate the unknown copula function. In order to use a few largest Fourier coefficients, the absolute value of the Fourier coefficients are arranged from largest to smallest and compare with $\sqrt{\Delta_n} = \sqrt{n^{-1}\log n \log m_n}$. By using the sample size of each data set $\sqrt{\Delta_n}$ is calculated, then according to the chosen algorithm of Fourier coefficients, these coefficients are obtained. With several start copula densities (Uniform, Gaussian, Clayton and, Frank), as it is shown in Tables 3 and 4, we have several estimations of copula density for rural and urban data sets.

For Urban data the sample size is $n = 18839$, with $\hat{c}_0(u,v) = 1$, calculation gives that $(\hat{\gamma}_{r_k s_k})^2 \ge \Delta_{18839}$ for $(r,s) = (1,1),(2,2),(3,3),(4,4)$, so $\hat{k} = 4$ and

$$\begin{aligned}\hat{c}^u(u,v) = 1 &+ 0.0764 b_1(u)b_1(v)\\ &+0.0592 b_2(u)b_2(v)\\ &+0.0429 b_3(u)b_3(v)\\ &+0.0313 b_4(u)b_4(v)\end{aligned} \tag{24}$$

With the Gaussian copula density as the start point, calculations give $(r,s) = (3,3),(4,4)$, so $\hat{k} = 2$ and

$$\begin{aligned}\hat{c}^G(u,v) = c_0(u,v;0.798)\\ + 0.0540 b_3(u)b_3(v)\\ + 0.0738 b_4(u)b_4(v)\end{aligned} \tag{25}$$

For the start with Frank copula density calculations give $(r,s) = (2,2),(3,3),(4,4)$, so $\hat{k} = 3$ and

$$\begin{aligned}\hat{c}^F(u,v) = c_0(u,v;6.42)\\ + 0.1203 b_2(u)b_2(v)\\ + 0.1804 b_3(u)b_3(v)\\ + 0.18212 b_4(u)b_4(v)\end{aligned} \tag{26}$$

For Rural data the sample size is $n = 19340$, with $\hat{c}_0(u,v) = 1$, calculation gives that $(\hat{\gamma}_{r_k s_k})^2 \ge \Delta_{19340}$ for $(r,s) = (1,1),(2,2),(3,3)$, so $\hat{k} = 3$ and

$$\begin{aligned}\hat{c}^u(u,v) = 1 &+ 0.0738 b_1(u)b_1(v)\\ &+ 0.0539 b_2(u)b_2(v)\\ &+ 0.0372 b_3(u)b_3(v)\end{aligned} \tag{27}$$

With the Gaussian copula density as a start point, calculations give $(r,s) = (2,2),(4,2)$, so $\hat{k} = 2$ and

$$\begin{aligned}\hat{c}^G(u,v) = c_0(u,v;0.812)\\ + 0.0412 b_2(u)b_2(v)\\ + 0.0352 b_4(u)b_2(v)\end{aligned} \tag{28}$$

For the start with Clayton copula density calculations give $(r,s) = (2,2)$, so $\hat{k} = 1$ and

$$\hat{c}^C(u,v) = c_0(u,v;2.067) + 0.699 b_2(u)b_2(v) \tag{29}$$

It should be noted that without using this method (selection method) among known copula densities, Frank copula and Clayton copula are the appropriate copulas for Urban and Rural data respectively, here these copulas can be chosen as starting points.

5.2. Investigating Performance of the Estimated Copula Function

To check the performance of the estimated copula densities, frequency of data is compared with the estimated probabilities, based on mean absolute relative error $(m.a.r.e)$, $|\frac{\hat{c}}{freq} - 1|$, on the same symmetric rectangles $(u = v = 0.25, 0.4)$ and asymmetric rectangles $(u = 0.25, v = 0.5; u = 0.5, v = 0.25)$ and also the corresponding upper tail rectangles.

It can be seen from Table 5 for Urban data copula density function $\hat{C}^G$ with Gaussian copula as a starting point has the least $m.a.r.e$. Also Table 6 shows that for Rural data copula density function $\hat{C}^G$ with Gaussian copula as the starting point has the least $m.a.r.e$.

Table 3. Results for Urban Data ($\sqrt{\Delta_{18839}} = 0.029$)

c_0	*Uniform*
$\hat{\theta}$	–
$\hat{\gamma}_{rs}$	$\hat{\gamma}_{11} = 0.0764$, $\hat{\gamma}_{22} = 0.0592$, $\hat{\gamma}_{33} = 0.0429$, $\hat{\gamma}_{44} = 0.0313$
$\hat{k}$	4
$\hat{c}^u$	$\hat{c}^u(u,v) = 1 + 0.0764 b_1(u)b_1(v) + 0.0592 b_2(u)b_2(v) + 0.0429 b_3(u)b_3(v) + 0.0313 b_4(u)b_4(v)$
c_0	*Gaussian*
$\hat{\theta}$	0.798
$\hat{\gamma}_{rs}$	$\hat{\gamma}_{33} = 0.0540$ $\hat{\gamma}_{44} = 0.0738$
$\hat{k}$	2
$\hat{c}^G$	$\hat{c}^G(u,v) = c_0(u,v;0.798) + 0.0540 b_3(u)b_3(v) + 0.0738 b_4(u)b_4(v)$
c_0	*Frank*
$\hat{\theta}$	1.01
$\hat{\gamma}_{rs}$	$\hat{\gamma}_{22} = 0.1203$ $\hat{\gamma}_{33} = 0.1804$ $\hat{\gamma}_{44} = 0.1821$
$\hat{k}$	3
$\hat{c}^F$	$\hat{c}^F(u,v) = c_0(u,v;6.42) + 0.1203 b_2(u)b_2(v) + 0.1804 b_3(u)b_3(v) + 0.18212 b_4(u)b_4(v)$

Table 4. Results for Rural Data ($\sqrt{\Delta_{19340}} = 0.0286$)

c_0	$Uniform$
$\hat{\theta}$ $\hat{\gamma}_{rs}$ $\hat{k}$ $\hat{c}^U$	$\hat{\gamma}_{11} = 0.0738 \quad \hat{\gamma}_{22} = 0.0539 \quad \hat{\gamma}_{33} = 0.0372$ 3 $\hat{c}^u = 1 + 0.0738 b_1(u)b_1(v) + 0.0539 b_2(u)b_2(v) + 0.0372 b_3(u)b_3(v)$
c_0	$Gaussian$
$\hat{\theta}$ $\hat{\gamma}_{rs}$ $\hat{k}$ $\hat{c}^G$	0.754 $\hat{\gamma}_{22} = 0.0412, \quad \hat{\gamma}_{42} = 0.0352$ 2 $\hat{c}^G(u,v) = c_0(u,v; 0.812) + 0.0412 b_2(u)b_2(v) + 0.0352 b_4(u)b_2(v)$
c_0	$Clayton$
$\hat{\theta}$ $\hat{\gamma}_{rs}$ $\hat{k}$ $\hat{c}^C$	2.067 $\hat{\gamma}_{22} = 0.0699$ 1 $\hat{c}^C(u,v) = c_0\,(u,v; 2.067) + 0.699 b_2(u)b_2(v)$

Table 5. The frequencies and approximations on different rectangles for Urban data

RECTANGLES	$freq$	$\hat{c}^U/freq$	$\hat{c}^G/freq$	$\hat{c}^F/freq$
$(0, 0.25) \times (0, 0.25)$	0.1726	0.986	0.980	1.022
$(0, 0.4) \times (0, 0.4)$	0.2985	0.985	0.997	1.009
$(0, 0.25) \times (0, 0.5)$	0.2290	0.991	0.991	1.001
$(0, 0.5) \times (0, 0.25)$	0.2296	1.090	1.001	1.004
$(0.75, 1) \times (0.75, 1)$	0.1652	1.044	1.015	1.019
$(0.6, 1) \times (0.6, 1)$	0.2950	1.009	1.004	1.021
$(0.75, 1) \times (0.5, 1)$	0.2286	0.992	1.002	1.012
$(0.5, 1) \times (0.75, 1)$	0.2247	1.003	1.013	1.008
$m.a.r.e.$		0.024	0.008	0.012

Table 6. The frequencies and approximations on different rectangles for Rural data

RECTANGLES	$freq$	$\hat{c}^U/freq$	$\hat{c}^G/freq$	$\hat{c}^C/freq$
$(0, 0.25) \times (0, 0.25)$	0.1739	0.917	0.981	1.058
$(0, 0.4) \times (0, 0.4)$	0.2959	0.988	0.994	1.042
$(0, 0.25) \times (0, 0.5)$	0.2271	0.995	0.994	1.031
$(0, 0.5) \times (0, 0.25)$	0.2303	0.987	0.977	1.052
$(0.75, 1) \times (0.75, 1)$	0.1516	0.979	1.008	0.982
$(0.6, 1) \times (0.6, 1)$	0.2859	1.008	1.019	0.985
$(0.75, 1) \times (0.5, 1)$	0.2223	1.021	1.013	0.992
$(0.5, 1) \times (0.75, 1)$	0.2199	1.014	1.024	0.968
$m.a.r.e.$		0.022	0.015	0.032

Table 7. The 0.99 and 0.95 estimated and real quantiles for Urban data

Probability	Quantile	Expected number	Real number
0.99	0.9654	189	182
0.95	0.9421	942	940

Table 8. The 0.99 and 0.95 estimated and real quantiles for Rural data

Probability	Quantile	Expected number	Real number
0.99	0.9791	194	190
0.95	0.8753	967	970

For more investigating of this new method, we considered the 99% (95%) quantile u^*, based on $\hat{c}^G$ (for both datasets) and the actual number of data points (u_i, v_i) outside the rectangle $(0, u^*) \times (0, u^*)$ with the expected number which is $(1 - 0.99)n$ or $(1 - 0.95)n$ was compared. The results are shown in Tables 7 and 8. In both tables, the expected numbers and the real numbers are close to each other.

6. Conclusions

In this paper in order to approximate copula density function for two variables, Income and Expenditure, of Iranian household, contamination families and selection models methods have been used. In this approach, a sequence of parametric copulas has been considered and in a few numbers of steps, accurate approximations for copula densities are obtained. By using the selection model method, the model complexity and number of model parameters have been balanced. It was shown that the best approximations for copula density function are the ones that are based on Gaussian starting copula. Also by using m.a.r.e. as a criterion, it has been shown that for both cases approximation with Gaussian copula as the starting point has the least mean absolute relative error.

REFERENCES

[1] Biau, G., Wegkamp, M., 2005. A Note on minimum distance estimation of copula densities. Statistics probability Letters 73, 105-114.

[2] Emberchts, P., Lindskog, F., McNeil, A., 2003. Modelling dependence with copulas and applications to risk managements. Rachev, S.T. (Ed.), Handbook of Heavy Tailed Distributions in Finance. Elsevier, Amsterdam, 329-384.

[3] Kallenberg, W.C.M., 2008. Modelling Dependence. Insurance: Mathematics and Economics, 2008, vol. 42, issue 1, 127-146.

[4] Kallenberg, W.C.M. 2009. Estimating copula densities, using model selection techniques. Journal of Insurance: Mathematics and Economics 45 209-223.

[5] McNeil, A., Frey, R., Emberchts, P., 2005. Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton.

[6] Nelsen, R.B., 1999. An Introduction to Copulas. Lecture Notes in Statistics, 139. Springer Verlag, New York.

[7] Sklar, A., 1959. Fonctions de repartition a n dimensions et leurs marges. Punl. Inst. Statist. Univ. Paris 8 229-231. 10.

[8] Sklar, A., 1996. Random variables, distribution functions, and copulas- a personal look backward and forward. In Distributions with Fixed Marginals and Related Topics (L. Ruschendorf, B. Schweizer and M.D. Taylor, eds). 1-14, Lecture notes monograph series 28, Institute of Mathematical 2 Statistics, Hayward, CA.